Humboldt-Universität zu Berlin

DISSERTATION

Data Accuracy in Bibliometric Data Sources and its Impact on Citation Matching

Zur Erlangung des akademischen Grades

Doctor philosophiae (Dr. phil.)

im Fach Bibliotheks- und Informationswissenschaft

eingereicht an der Philosophischen Fakultät I

von

Mag. (FH) Marlies Olensky

Präsident der Humboldt-Universität zu Berlin: Prof. Dr. Jan-Hendrik Olbertz

Dekan der Philosphischen Fakultät I: Prof. Michael Seadle, Ph.D.

Gutachter/in: 1. Prof. Vivien Petras, Ph.D.

2. Prof. Birger Larsen, Ph.D.

Datum der Einreichung: 17. Oktober 2014

Datum der Disputation: 17. Dezember 2014

ABSTRACT

Data Accuracy in Bibliometric Data Sources and its Impact on Citation Matching
by Marlies Olensky

Is citation analysis an adequate tool for research evaluation? This complex question can be addressed from a variety of angles. At the core of this issue stands the question whether the underlying citation data is sufficiently accurate to provide meaningful results of the analyses and if not, whether the citation matching process can rectify inaccurate citation data. Thus, this doctoral research tackles the question from a data analysis point of view. It investigates the accuracy of bibliographic data in bibliometric data sources, that is used in citation analyses.

In this research, inaccuracies in bibliometric data sources are defined as discrepancies in the data values of bibliographic references, since they are the essential part in the citation matching process and, therefore, have the greatest impact on their accuracy. A stratified, purposeful data sample was selected to examine typical cases of publications in Web of Science (WoS). The bibliographic data of 3,929 references was assessed in a qualitative content analysis to identify prevailing inaccuracies in bibliographic references that can interfere with the citation matching process. The inaccuracies were analyzed, categorized and organized into a taxonomy. Additionally, their frequency was studied to determine any strata-specific patterns, i.e. whether, for example, certain document types or languages are more prone to contain more or different kinds of inaccuracies. To pinpoint the types of inaccuracies that influence the citation matching process, a specific subset of citations was investigated. The subset consisted of citations not successfully matched by WoS, but identified manually in its *Cited Reference Search*, i.e. missed citations. The results were triangulated with five other data sources: with data from two bibliographic databases in their role as citation indexes (Scopus and Google Scholar) and with data from three applied bibliometric research groups (CWTS, iFQ and Science-Metrix).

In total, 5.57% missed citations were identified in the *Cited Reference Search* of WoS. In the citations missed by WoS, 57% of inaccuracies were caused by authors, 12% were due to the

citation style which WoS did not process correctly, and 31% of inaccuracies were traced back to the data handling process or to inaccurate data that had originally been supplied to WoS. The matching algorithms of CWTS and iFQ were able to match around two thirds of these citations correctly. Scopus and Google Scholar also handled more than 60% successfully in their matching. Science-Metrix only matched a small number of references (5%) due to the fact that it usually incorporates the article title provided in the Scopus raw citation data in its citation matching process. While some inaccuracies have more impact on the citation matching process than others, completely incorrect starting page numbers and transposed publication years can cause a citation to be missed in all data sources. However, more often it is a combination of more than one kind of inaccuracy in more than one field that leads to a non-match. Based on these results, proposals are formulated that could improve the citation matching processes of the different data sources. They build on the inclusion of as many bibliographic fields as possible and of variation thresholds for the data values to be matched.

ZUSAMMENFASSUNG

Data Accuracy in Bibliometric Data Sources and its Impact on Citation Matching
von Marlies Olensky

Ist die Zitationsanalyse ein geeignetes Instrument zur Forschungsevaluation? Diese komplexe Frage kann aus einer Vielzahl von Blickwinkeln beleuchtet werden. Im Kern steht vor allem die Frage, ob die zugrunde liegenden Zitationsdaten ausreichend fehlerfrei sind, um aussagekräftige Ergebnisse der Analysen zu erzielen, beziehungsweise sollte dies nicht der Fall sein, ob der Prozess, der die zitierenden und zitierten Artikel einander zurordnet, ausreichend robust gegenüber Ungenauigkeiten in den Daten ist. Diese Dissertation beschäftigt sich daher mit der Analyse der Richtigkeit von bibliographischen Daten aus bibliometrischen Datenquellen, die zur Zitationsanalyse herangezogen werden.

Ungenauigkeiten in bibliometrischen Datenquellen werden als Unterschiede in den Datenwerten der bibliographischen Angaben definiert, da diese den Prozess der Zuordnung von zitierenden zu zitierten Artikeln wesentlich beeinflussen und größte Auswirkung auf dessen Genauigkeit haben. Die untersuchten Daten setzen sich aus gezielt ausgewählten Publikationen des Web of Science (WoS) zusammen, welche eine geschichtete Stichprobe ergeben. Die bibliographischen Daten von 3.929 Referenzen wurden in einer qualitativen Inhaltsanalyse bewertet, um die Verteilung von Ungenauigkeiten in Literaturangaben, die den Zuordnungsprozess von Zitationen behindern könnten, zu bestimmen. Die bibliographischen Ungenauigkeiten wurden zusätzlich in einer Taxonomie zusammengefasst. Außerdem wurden die verschiedenen Schichten der Stichprobe auf auftretende Muster von Ungenauigkeiten untersucht, um zum Beispiel herauszufinden, ob bestimmte Dokumenttypen oder Sprachen mehr Ungenauigkeiten beziehungsweise verschiedene Arten von Ungenauigkeiten aufweisen. Um genau festzulegen, welche von diesen tatsächlich den Zuordnungsprozess von Zitationen beeinflussen, wurde eine spezifische Untergruppe von Zitationen untersucht. Diese Teilmenge bestand aus Referenzen, die von WoS nicht erfolgreich dem jeweilig zitierten Artikel zugeordnet wurden, aber in der *Cited Reference Search* identfiziert werden konnten, sogenannte fehlende Zitierungen (*missed citations*). Die Ergebnisse wurden mit den Daten zweier weiterer

bibliographischen Datenbanken, Scopus und Google Scholar, sowie den Daten dreier angewandter bibliometrischer Forschungsgruppen, CWTS, iFQ und Science-Metrix, trianguliert.

Im Ergebnis wurden insgesamt 5,57% fehlende Zitierungen in WoS identifiziert. In diesen wurden 57% der Ungenauigkeiten von den zitierenden Autoren verursacht; 12% entstanden aufgrund des Zitierstils, der in WoS nicht richtig verarbeitet wurde; die restlichen 31% der Ungenauigkeiten sind auf den Datenverarbeitungsprozess beziehungsweise auf Daten, die bereits fehlerbehaftet an WoS geliefert wurden, zurückzuführen. Die Zuordnungsalgorithmen von CWTS und iFQ konnten rund zwei Drittel dieser Zitierungen erfolgreich in ihren Datenbanken zuordnen. Scopus und Google Scholar konnten ebenso über 60% der fehlenden Zitierungen erfolgreich mit dem entsprechenden zitierten Artikel verbinden. Science-Metrix war es nur möglich eine geringe Anzahl an Referenzen (5%) dem richtigen zitierten Artikel zuzuordnen, da diese Forschungsgruppe in der Regel den Artikeltitel, der Teil der Zitationsrohdaten in Scopus ist, in den Zuordnungsprozess miteinbezieht. Während einige Ungenauigkeiten mehr Einfluss auf den Zuordnungsprozess von Zitationen haben als andere, können vollkommen falsche erste Seitenzahlen sowie Zahlendreher in Publikationsjahren in allen Datenquellen nicht richtig zugeordnete Zitierungen verursachen. Häufig ist es jedoch eine Kombination von mehreren Arten von Ungenauigkeiten in mehr als einem bibliographischen Datenfeld, die eine korrekte Zuordnung verhindern. Basierend auf diesen Ergebnissen wurden Lösungsvorschläge formuliert, die im Stande sind den Zuordnungsprozess von Zitationen in bibliometrischen Datenquellen zu verbessern. Im Fokus liegt die Einbeziehung möglichst vieler Datenfelder, sowie variabler Schwellenwerte für die zuzuordnenden Datenwerte aus bibliographischen Referenzen.

ACKNOWLEDGEMENTS

First of all I would like to thank my fantastic Doktormutter, Vivien Petras, who kindly adopted me after my original Doktorvater, Stefan Gradmann left our institute for a position at the KU Leuven. Her drive, enthusiasm, inspirational ideas and critical questions helped shape the details of this research. Unfortunately, Stefan Gradmann was not able to continue his support for this dissertation, but I will be forever grateful to him for believing in me and assigning me a position at Humboldt involving the conferral of a doctorate. I particularly wish to extend my gratitude to Birger Larsen, who, after presenting my first proposal for this dissertation at the joint doctoral colloquium with the Royal School of Library and Information Science in Copenhagen, happily agreed to be my second advisor. Without his insights and expert knowledge on bibliometrics this dissertation would not have been feasible. Thanks to both of you for always asking the right questions.

My special thanks go to the three applied bibliometric research groups, iFQ, CWTS and Science-Metrix, who, despite their busy schedules, made time to support my research project and kindly provided data for my analysis. In particular, I would like to thank Stefan Hornbostel of iFQ and Paul Wouters of CWTS who initiated the collaboration; Eric Archambault and Philippe Deschamps of Science-Metrix, Nees Jan van Eck of CWTS and Marion Schmidt of iFQ who not only provided their data, but also patiently answered all my questions. I am also grateful to have found two such reliable, fast-working and enthusiastic student assistants, Till Erhart and Apoorva Rajiv – the data collection is your work.

In her advice on how to write a dissertation, Joan Bolker recommends NOT moving across oceans in the year you are about to finish your project. I did it anyway and my experience at the National Taiwan University (NTU) would not have been the same without a lot of people: my uncle Walter, Prof. Clarence Chu, Prof. Mu-Hsuan Huang, Prof. Shanju Lin, Yilin, Sunny, Ying Ta and Caitlin. And a big thank you goes to my foreign NTU crowd for keeping me socially sane: Maryline, Alex, Joris, Seb, Jin, Sylvain, and especially Maud who additionally was kind (and brave) enough to help with checking my manuscript.

I am very thankful to Jenny, my very first PhD buddy, who not only gave me the initial idea of this topic, but continuingly listened to my ups and downs. Not enough thanks can be given to my colleagues – turned PhD buddies – turned friends, Maria and Juliane. Life in Berlin would not have been the same without you. A big thank you to all three of you for reading parts of the manuscript and your valuable feedback. A special thanks also goes to my editor, Carol Marshall, who worked her magic on my words and just always knew how to give them the finishing touch.

Ein großes Dankeschön geht an Priska für ihre unermüdliche Unterstützung beim Korrekturlesen verschiedenster Paper, Proposals und Kapiteln. Und danke auch an Julia, die sich tapfer durch die Methoden- und Ergebniskapitel gequält hat. Mein besonderer Dank gilt meinen Eltern für ihre bedingungslose und fortwährende Unterstützung, auch wenn dies bedeutet nicht in derselben Zeitzone zu leben. Mein größter Dank gilt meinem Ein und Alles, Marcus, der immer genau das war, was ich gerade am meisten brauchte: Coach, Cheerleader oder auch Diktator. Danke für das gemeinsame Durchhalten und dafür, dass du mich täglich zum Lachen bringst.

Table of Contents

List of Figures

List of Tables

Abbreviations

AHCI	Arts & Humanities Citation Index
ASCII	American Standard Code for Information Interchange
CWTS	Centre for Science and Technology Studies, Leiden
DQ	Data Quality
DOI	Digital Object Identifier
ERA	Excellence in Research of Australia
FRBR	Functional Requirements for Bibliographic Records
GS	Google Scholar
IAC	Inaccuracy code
iFQ	Institut für Forschungsqualität, Berlin
ISI	Institute for Scientific Information
ISO	International Organization for Standardization
JCR	Journal Citation Report
JI	WoS field tag – ISO source abbreviation of publication name
JIF	Journal Impact Factor
LDF	Levenshtein Distance Function
LIS	Library and Information Science
MCR	Missed Citation Rate
NS	Natural Sciences
OCR	Optical Character Recognition

OPED	OPposite the EDitorial page
REF	Research Excellence Framework (UK)
SCI	Science Citation Index
SCIE	Science Citation Index Expanded
SIGMETRICS	Special Interest Group on Measurement and Evaluation
SM	Science-Metrix, Montreal
SO	WoS field tag – full publication name
SSCI	Social Sciences Citation Index
SSH	Social Sciences & Humanities
WoS	Web of Science
WWW	World Wide Web

1 INTRODUCTION

1.1 Research problem

Research evaluation is becoming increasingly important. Particularly in the current era of the knowledge economy, it is essential to remain internationally competitive (D'Angelo, Giuffrida & Abramo, 2011). Hence, universities and research institutes, as the central hub of the knowledge production of a nation, are subjected to an evaluation of their output. The quality of research is usually determined by peer review or by indicators, often referred to as science indicators, based on publication and citation statistics which measure its productivity and impact (Leydesdorff, 2008). Since citing another researcher's work is a sign of reproducing, corroborating and supporting and sometimes even refuting the ideas and results of the researcher, citations express the impact of a researcher's work (Bornmann & Marx, 2013). National research evaluation initiatives, such as the Excellence in Research of Australia (ERA), employ peer review as well as citation analysis, others rely on peer review only (e.g. the Research Evaluation Framework (REF) assessing UK higher education institutions). Especially in the social sciences and humanities (SSH), research assessment for governments is conducted via peer review, which is time-consuming and can also be biased (Kousha & Thelwall, 2007). Other research evaluations, such as the rankings of universities (e.g. the Shanghai Ranking[1] or the Leiden Ranking[2]), use bibliometric indicators in addition to other metrics, such as the number of alumni, to assess the research excellence of universities. Citation profiles of researchers are often consulted in connection with decisions pertaining to recruitment and career advancement in publicly funded research organizations (Steele, Butler & Kingsley, 2006; D'Angelo et al., 2011) or the assignment of research funds (Meho & Yang, 2007). Additionally, citation counts and rankings can help identify subject experts, who are then employed to review applications, manuscripts and project results (Meho & Yang, 2007).

[1] http://www.shanghairanking.com/
[2] http://www.leidenranking.com/

While some research evaluation initiatives are skeptical whether citation analysis is an adequate means to measure scholarly output (Research Evaluation Framework, 2014), it is, next to peer review, the best alternative available, since it is faster, completely objective and impartial compared to peer review. However, only if certain standards are met, can bibliometric indicators be used in research evaluation: it is necessary to ensure that the bibliographic data is complete (Moed, Burger, Frankfort & van Raan, 1985), the methodology and data processing are adequate and documented (Moed et al., 1985), the data sources are described (Glänzel, 1996) and bibliometric indicators are exactly defined (Glänzel, 1996) as well as correctly calculated and handled (Bornmann, Mutz, Neuhaus & Daniel, 2008). Hence, many different factors contribute to the accurate results of citation analyses. However, one of the most important resources is the data itself – the citations.

The data is usually provided by bibliographic databases which also store citing references and, therefore are referred to as citation indexes. According to their use as a data source for bibliometric analyses, these databases can also be labeled as bibliometric data sources. Web of Science[3] (WoS) by Thomson Reuters, Scopus[4] by Elsevier, and Google Scholar[5] (GS) are the three big players on the market. Hence, the data quality as well as the correct matching of citations in these databases play an important role in citation analysis. Although the error correction process of references in WoS is not a trivial matter (Jacsó, 2004), bibliometric data sources still contain missed citations, i.e. stray references, and inaccuracies in their data (van Raan, 2005). Experts in bibliometrics warn against using data from bibliometric data sources blindly for citation analysis and even imply that one should not rely implicitly on the results of bibliometric indicators calculated by a citation index, such as WoS (Reedijk, 1998; Moed, 2002), especially in the case of less cited articles (Kostoff, 2002). Hence, as long as database providers, such as WoS, have not "implement[ed] a procedure of systematically identifying and correcting erroneous source or citation data on a paper-by-paper basis" (Reedijk, 1998, p. 769), data quality problems in bibliometric data sources are far from being solved (Franceschini, Maisano & Mastrogiacomo, 2013b). Even though the responsibility for accurate references initially lies with authors, editors and publishers (Garfield, 1983; 1990), ultimately, in particular when citation counts and citation analyses based on them are provided, the responsibility lies "(at least morally) for the quality, or lack thereof, of their content" with database publishers (Tenopir, 1995, p. 124).

[3] http://wokinfo.com/
[4] http://www.scopus.com/
[5] http://scholar.google.com/

Questions concerning the comparability of bibliometric indicators, especially across disciplines, as well as the lack of standardization of the calculation processes have driven the discussions in the bibliometric community for years (van Raan, 2005). In their studies, researchers address the development and evaluation of bibliometric indicators in citation analyses and endeavor to determine which database (e.g. WoS, Scopus, GS, etc.) is the most appropriate source for their analyses. To decide on "the" data source for citation analysis, characteristics of citation indexes, such as the coverage and language of journals, the selection process of journals as well as the overlap of documents and citation counts between sources, have been the subject of several studies (e.g. Archambault, Campbell, Gingras & Larivière, 2009; Meho & Yang, 2007). While these factors are without doubt important decision criteria, an even more substantial aspect, namely the underlying citation data and the process by which it is matched, has not been investigated in-depth. Very few authors (e.g. Hildebrandt & Larsen, 2008; Larsen, Hytteballe Ibanez & Bolling, 2007; Moed, 2005) have studied data accuracy in bibliometric data sources before, and none of them with the goal of finding a standardized categorization of inaccuracies and/or determining their impact on different citation matching algorithms. They report on missed citation rates between 5 and 12% in WoS. It could be argued that in bibliometric studies on the macro-level the missed citation rates up to 12% may not influence the ranking of universities or countries. However, on the level of individual researcher assessment they are far more likely to impact the ranking of researchers (Garfield, 2005), but no substantiated research has investigated this issue so far.

In their citation matching processes, the databases use matching algorithms that are not publicly available because of competitive advantage. In recent years, sophisticated algorithms for matching cited and citing articles have been developed by applied bibliometric research groups (Neuhaus & Daniel, 2008) operating on the raw citation data provided by WoS or Scopus, which should rectify incorrect data in references. Only one applied bibliometric research group, the Institut für Forschungsqualität in Berlin, revealed parts of the research process of developing such a matching algorithm (Schmidt, 2012); apart from this exception, no published research exists to date. Hence, no study has ever evaluated the efficiency of the algorithms or compared them with each other. Nevertheless, experts in the bibliometric community stress the need for standard match keys in order to achieve comparability of bibliometric studies (Glänzel, 1996) and even suggest that citation indexes need to be rebuilt into a new system that "[…] is accurately [sic] enough to use it for the calculation of bibliometric indicators and to apply it for evaluation purposes" (van Raan, 2005).

1.2 Research questions

This doctoral research contributes to increasing the transparency of citation analysis results in order to use them as a research assessment tool by investigating how well citation matching algorithms handle inaccurate data. It aims to convey a full understanding of the characteristics, patterns and causes of inaccurate bibliographic data that can influence the citation matching process. Therefore, it provides unprecedented analyses of the handling of inaccuracies in bibliographic references in the citation matching process. Data from the three major bibliometric data sources, WoS, Scopus and GS, are compared as well as data kindly provided by three leading applied bibliometric research groups, Centre for Science and Technology Studies in Leiden (CWTS)[6], Institut für Forschungsqualität in Berlin (iFQ)[7] and Science-Metrix[8] (SM) in Montreal. The analysis identifies inaccuracy patterns in bibliographic references and reveals the types of inaccuracies the databases themselves and those of the applied research groups are able to rectify in their citation matching algorithms and which of the inaccuracies lead to non-matched citations, i.e. lost or missed citations, that are, therefore, not considered in citation analyses. Based on the findings, proposals are put forward to optimize citation matching algorithms, reduce the number of non-matched citations and draw a more accurate picture of citation profiles.

The following research questions are addressed which describe a stepwise research process in which one research question incorporates the results of the preceding ones:

RQ1 What types of inaccuracies occur in bibliographic data?
- How can they be categorized?
- How frequent is their incidence in bibliometric data sources?
- Can patterns be identified?

RQ2 What types of inaccuracies cause missed citations?
- How well do citation matching algorithms handle inaccurate data?

RQ3 How can the number of non-matches in the citation matching process be reduced?

The results are threefold. First, a taxonomy of bibliographic inaccuracies is developed which helps to reveal whether inaccuracy patterns in bibliographic references can be identified that

[6] http://www.cwts.nl/
[7] http://www.forschungsinfo.de/
[8] http://www.science-metrix.com/

can be translated into machine-readable rules for data matching. Second, the analysis sheds light on the inaccuracy categories that lead to missed citations in the bibliometric data source WoS and then triangulates this result with the other five bibliometric data sources, Scopus, GS, CWTS, iFQ and Science-Metrix, in order to obtain more valid results. Third, the dissertation formulates proposals as to how the citation matching process could be improved. To establish to what extent missed citations influence the result of bibliometric calculations is beyond the scope of this research.

The unique contribution of this dissertation is the systematic investigation of inaccuracies in citations, which is the first of its kind. Moreover, the citation matching algorithms of three leading applied bibliometric research groups have never been published or compared with each other. In this doctoral research we not only investigate the differences in the data of the three main bibliometric data sources, which are available to every subscriber and researcher, but we were in the privileged position of having access to matched citation data from all three applied bibliometric research groups and were thus able to evaluate them. Therefore, all bibliometric data sources investigated could benefit from this study, as it could trigger changes in their customized matching algorithms. In particular, the applied bibliometric research groups can benefit from the data corpus created during this research, consisting of manually checked citations, i.e. both missed citations and false positives are verified, which provides an ideal opportunity for them to use it for further experiments with their matching algorithms. Hence, the dissertation not only theoretically contributes to the research of increasing the transparency of results of citation analyses, but could have a direct, practical impact on the bibliometric studies carried out by the three institutions (e.g. CWTS Leiden Ranking). In a nutshell, this doctoral research provides unique findings that have the potential to influence the entire bibliometric research community.

Moreover, laymen, i.e. scientists who are obliged to prepare their own citation or impact profiles, or librarians, who are often employed to carry out citation analyses for universities or research institutions, can also benefit from the results. They will receive an indication of how reliably the citation matching in each data source works and how much manual effort has to be invested when evaluating researchers and their citations. Additionally, it informs all scientific authors about the parts of references that require special attention in order to provide an as accurate a basis as possible for citation matching and analysis.

1.3 Organization of the dissertation

The dissertation is organized as follows: Chapter 2 discusses important bibliometric terminology used in citation analysis and bibliometric studies. In particular, the process of citation matching and the concept of a missed citation are explained. The chapter also defines the term bibliometric data source as used in this research. Chapter 3 specifies the context of data accuracy within the data quality literature and justifies the focus of this research on data values. It further elaborates on the characteristics of bibliographic data accuracy and reviews how it has been assessed in previous studies. Data inaccuracy is defined as understood in this research. Chapter 4 explains how data accuracy in bibliometric data sources can be defined and presents the current state of research on inaccuracies in bibliographic data values. Chapter 5 discusses the methodology employed in this doctoral research. A qualitative content analysis is applied, adapted to the characteristics of bibliographic data. The chapter also presents the selection process of a multifaceted data sample and reports on the process of data collection. Chapter 6 presents the results of the qualitative content analysis of inaccuracies, i.e. a coding scheme for bibliographic inaccuracies, and organizes them into a taxonomy. Chapter 7 introduces the results of the quantitative analysis of bibliographic inaccuracies. The overall occurrences of inaccuracies are discussed as well as specifics of the different facets of the data sample. Chapter 8 focuses on the evaluation of missed citations. The distribution of missed citations in WoS as well as the ability of the other data sources (Scopus, GS, CWTS, iFQ and Science-Metrix) to match them are examined. Chapter 9 introduces proposals to improve citation matching based on the empirical findings described in Chapters 6 to 8. Chapter 10 concludes this dissertation by giving an overview of its contribution as well as an outlook on future work.

2 Defining Bibliometric Terminology

The theoretical part of this dissertation begins with an overview of the bibliometric terminology used in this research. First, the terms bibliometrics and citation analysis are introduced. In citation analysis (section 2.2), we explain the concepts of cited and citing articles, citation window, citation matching and missed citation. Bibliometric indicators are the results of citation analysis and are presented in section 2.3. The concept of a bibliometric data source as understood in this dissertation is defined in section 2.4. The chapter concludes with a summary in section 2.5.

2.1 Bibliometrics

Library and information science (LIS) and related fields (e.g. Science and Technology Studies) have developed sets of methodologies that allow the measurement of the production, use, re-use and dissemination of different kinds of information (Björneborn & Ingwersen, 2004). These sets of methodologies developed into research subfields, which are illustrated in Figure 1. Informetrics is the superordinate term (Tague-Sutcliffe, 1992). Bibliometrics and scientometrics evolved from the same idea, which was to analyze citations. In the 1960s, Eugene Garfield laid the foundation for citation analysis with his invention of the Science Citation Index (SCI) (cf. section 2.4.2). Later, access to online citation databases opened up a wide range of possibilities to study citations. In particular, the development of scientific domains, including growth, specialization, collaboration, impact, and obsolescence of literature and concepts, can be studied (Björneborn & Ingwersen, 2004). Scientometrics denotes the investigation of a researcher's publishing performance. In scientometrics, mainly scientific publications and citations are quantitatively and statistically analyzed. In contrast, bibliometrics is not limited to the study of scientific publications and is, therefore, used as a superordinate concept. Bibliometrics includes the study of bibliometric distribution, citation

analysis, library use, co-citation analysis, co-word analysis, and bibliographic coupling. Cybermetrics and webometrics are additional research fields which evaluate output published on the World Wide Web (WWW).

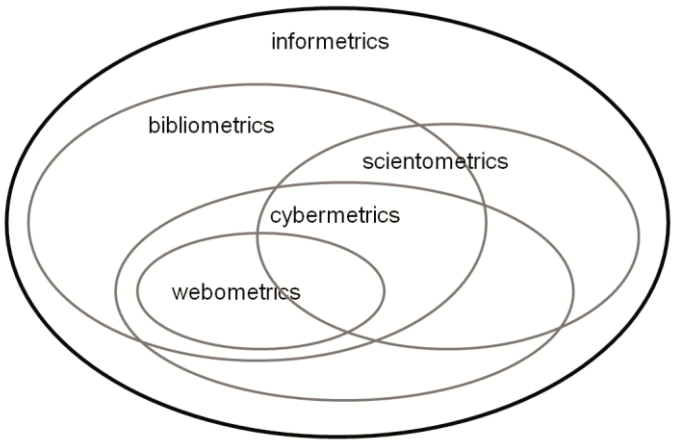


Figure 1: Relationships between the LIS fields of informetrics, bibliometrics, scientometrics, cybermetrics and webometrics[9] (Björneborn & Ingwersen, 2004)

In recent years the amount of scientific output has increased immensely (Priem, Taraborelli, Groth & Neylon, 2010). With a shift in publication behavior towards the WWW, the traditional means of conceiving and filtering out important research results, such as peer review or citation analysis, are complemented by another form of metrics based on the impact on the Social Web: altmetrics (Priem et al., 2010). Even though this new metric is an important advancement to capture scientific output on the web, it is still in an early stage of development and is compared with the results of citation analysis (e.g. Zahedi, Costas & Wouters, 2014).

For the time being, citation analysis is still the most important informetric element in research evaluation (e.g. the CWTS Leiden ranking or the Shanghai ranking of universities worldwide). Consequently, we focus on citation analysis and its components in this doctoral research.

2.2 Citation analysis

Citation analysis is one of the methods out of the bibliometric and scientometric toolbox which investigates, inter alia, the number of publications, the number of citations received as well as a number of bibliometric indicators that are calculated on the basis of these counts. In

[9] The sizes of the overlapping ellipses are for the sake of clarity only.

this section, we discuss the concepts of cited and citing articles, citation matching and missed citation.

2.2.1 Cited and citing articles

A cited article is one that has been referenced by one or more articles. An article citing another article is called a citing article and holds a reference to one or more cited articles. A reference can also be referred to as a citation, citing reference or cited reference. Cited articles can also be designated as target articles, because they are the target to which citing articles are matched. Another term for citing articles, therefore, is source articles, as they are the source of the citation matching process (Moed, 2005; van Raan, 2005). Buchanan (2006) uses the definitions conversely and refers to cited articles as source articles and citing articles as target-source articles. We think that this definition complicates the issue and, therefore, adhere to the definition used by Moed (2005) and van Raan (2005). Figure 2 gives an example of target and source articles that cite each other. The arrows show the citation direction from the source to the target article. Two of these articles are at the same time target articles, i.e. cited, and source articles, i.e. citing. The references cited in an article are available in citation indexes as cited reference information (cf. section 2.2.4).

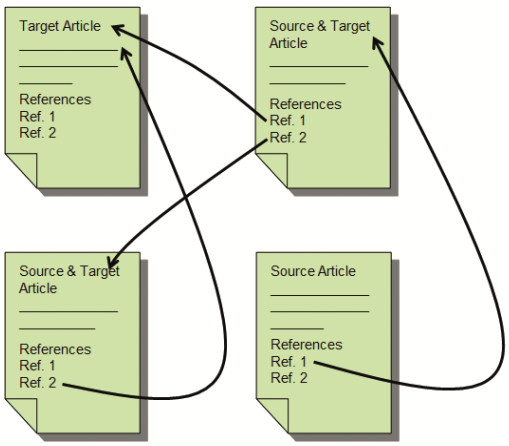
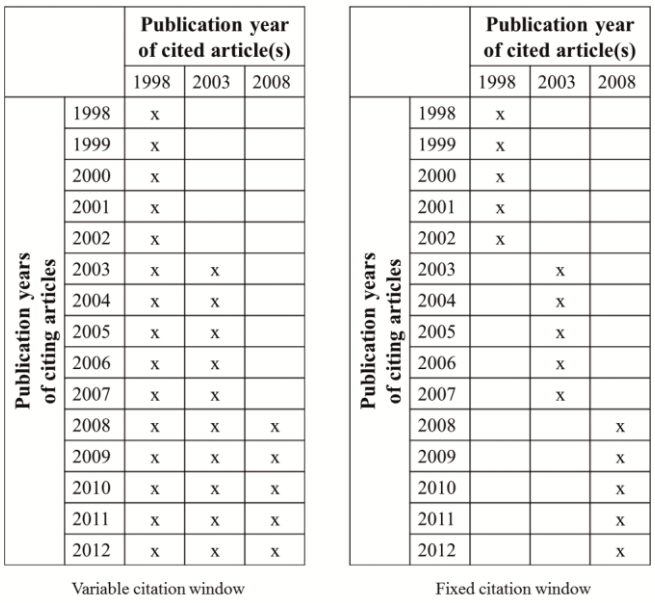


Figure 2: Target and source articles

2.2.2 Citation window

A citation window is the period of time allowed for publications to gather citations. A citation window in a citation analysis can either be variable or fixed. A variable citation window accumulates citations over the years, starting with the publication year of the document. Yet,

the end of the citation window is the same for all documents investigated. An example of a fixed citation window is a five-year citation window that would only consider citations accumulated during the first five years following the publication of a document. Figure 3 illustrates both variants. In citation analyses, a variable citation window is usually used when a large data sample is needed and the comparison of citation rates is not the focus. For instance, it can be used to measure collaboration between researchers (Levitt & Thelwall, 2009). A fixed citation window ensures that the citation rate has less variation (Katz, 1999), since on average a document's citation count increases and peaks in the third and fourth year after its publication; afterwards, the citation rate decreases until it has received about 80% of the total number of citations after about eight years after publication (Narin, 1976). Therefore, a fixed citation window can be applied to compare citation rates of documents.



		Publication year of cited article(s)		
		1998	2003	2008
Publication years of citing articles	1998	x		
	1999	x		
	2000	x		
	2001	x		
	2002	x		
	2003	x	x	
	2004	x	x	
	2005	x	x	
	2006	x	x	
	2007	x	x	
	2008	x	x	x
	2009	x	x	x
	2010	x	x	x
	2011	x	x	x
	2012	x	x	x

Variable citation window

		Publication year of cited article(s)		
		1998	2003	2008
Publication years of citing articles	1998	x		
	1999	x		
	2000	x		
	2001	x		
	2002	x		
	2003		x	
	2004		x	
	2005		x	
	2006		x	
	2007		x	
	2008			x
	2009			x
	2010			x
	2011			x
	2012			x

Fixed citation window

Figure 3: Variable vs. fixed citation window

2.2.3 Citation matching

Citation matching is the process that matches a citing reference in an article to its cited article. Based on citation matching, indicators that measure the impact of an article can be calculated (cf. section 2.3). The reliability of these indicators "strongly depends on the accuracy with which citation links are identified" (Moed, 2005, p. 173). The accuracy of the citation links is in turn influenced by the accuracy of the references in the citing articles and the accuracy with which the bibliographic data is extracted and handled by citation indexes, such as WoS or Scopus.

The bibliographic data employed in citation matching was influenced, maybe even determined, by the first available source of bibliographic and citation data: the Web of Science. At the beginning of the SCI, the Institute for Scientific Information (ISI) decided to extract the following information from the bibliography of an article to use in its citation matching process and has not changed the originally selected fields since: first author, source title (= publication name), year, volume number and starting page (Moed, 2005). Due to the high cost of storage at that time, ISI had opted to only cover the first author from a citing reference and, therefore, had been able to provide greater coverage of source titles (Garfield, 1990). Even though mass storage has become cheaper in the past few decades, Thomson Reuters has not changed its policy for extracting citing references. Nowadays, a database producer usually obtains bibliographical data electronically, directly from the publisher, which is the case with many journals processed by Thomson Reuters (Moed, 2005).

The actual process of citation matching involves so-called *match keys*, which consist of a combination of the above-mentioned bibliographic fields to uniquely match citing references to the correct cited articles. In this sense, citation matching processes are deterministic models of record linkage, as they lead either to a match or non-match of target and source articles (Synnestvedt, 2007). Match keys are varied in different steps of the matching process (Moed & Vriens, 1989; Synnestvedt, 2007; Schmidt, 2012) and in each step unique matches are extracted. The remaining unlinked articles form the input for the next step of the matching process, which continues with a different match key, i.e. set of bibliographic fields (Synnestvedt, 2007; Schmidt, 2012; P. Deschamps, personal communication, February 25, 2014).

The first reported "special search key", which was intended to characterize each publication uniquely in an evaluation of publishing performance and citation impact, consisted of "the first four letters of the author name, the last two digits of the publication year, the first character of journal title, journal volume and starting page number" (Braun, Glänzel & Schubert, 1985, pp. 406-407). Another study by Yannakoudakis, Ayres & Huggill (1990) matched citations from seven different (at that time non-standardized) databases and employed a basic match key consisting of author names and article titles to match the records. They used the eight least frequent digits or letters from the original article title or the translation as well as the eight least frequent from the first author's surname or a corporate body that was identified as the author. The main problems they encountered were due to transcribed and translated article titles from other languages into English. However, 45% of the citations were linked precisely. Apart from this, the rare occasions when researchers implicitly discussed the match keys they

had employed are in studies of inaccuracies in citation indexes (cf. section 4.2). The only documented citation matching algorithm was published by iFQ (Schmidt, 2012). It employs 40 different match keys, the Damerau-Levenshtein distance function and it allows for combinations of up to four wrong bibliographic fields (Schmidt, 2012). This citation matching algorithm is not yet used in the production process, but is still in development (M. Schmidt, personal communication, August 10, 2014). Other applied bibliometric research groups refrain from publishing the details of their citation matching algorithms to protect their competitive advantage (cf. section 2.4.3).

Apart from the Damerau-Levenshtein distance function, other fuzzy string matching methodologies can be employed in citation matching algorithms. For example, Abdulhayoglu & Thijs (2013) present an approach to match publication lists to WoS and Scopus records. They use *n*-grams based on the Levenshtein distance score for one entire record. They calculate several similarity scores and use them as variables in a kernel discriminant analysis. When adjusting the parameters they observed a trade-off between false positive and false negative matches. Christen (2006) carried out experiments comparing pattern matching algorithms for author names. The results revealed that there is no single best technique and that similarity measure calculations can have dramatic effects on the matching quality (Christen, 2006). He recommends data parsing (eliminating space characters and punctuation marks) and if one knows that the data contains many nicknames, a dictionary-based, name standardization should be applied before the matching process (Christen, 2006). Names that were parsed into separate fields can best be assessed by the Jaro-Winkler string comparator, which performs well for both given and last names. The longest common sub-string technique is suitable for unparsed names which may contain swapped strings (Christen, 2006). Performance-wise he reports that phonetic matching is a faster method (Christen, 2006). String matching methodologies and algorithms that could potentially be applied in citation matching are the following (Christen, 2006):

- Relative Levenshtein distance: relates the edit distance to the length of the assessed value.
- Damerau-Levenshtein distance: counts a transposition of two characters as only one edit.
- Bag distance: compares the single characters of each string in a pre-defined bag and disregards the order. It is a good means to filter out candidate matches.

- Smith-Waterman: this algorithm was developed for DNA matching. It works similar to an edit distance, but allows for gaps and character-specific match scores (e.g. similar sounding characters could be assigned a higher match score).
- Longest common substring: this algorithm finds, and repeatedly eliminates, the longest common sub-string (up to a minimum length, which is usually 2 or 3) of two strings that are to be matched. The resulting scores are used for calculating an edit distance.
- *n*-grams or *q*-grams: are sub-strings of length *q* in longer strings. Commonly used *n*-grams are unigrams (*n* = 1), bigrams (*n* = 2, also called digrams) and trigrams (*n* = 3). For example, 'peter' contains the bigrams 'pe', 'et', 'te' and 'er'. A similarity is calculated based on the overlap of the *n*-grams. *n*-grams are specifically useful in detecting and correcting typographical errors in bibliographic databases (O'Neill & Vizine-Goetz, 1988).
- Variations of the *q*-grams: positional *q*-grams (that also compare the position of the *q*-gram in the string), skip-grams (that also compare *q*-grams made by skipping a character in between).
- Sorted Winkler: if the value consists of more than one string, the strings are first ordered alphabetically. Therefore, a jumbled order of strings in the article title or author names (unless they only contain initials) would not be considered as a discrepancy.
- Permuted Winkler: all kinds of possible permutations of words are performed and the maximum of all similarity values calculated is returned.

These string matching methodologies are useful tools to overcome inaccuracies in data values and match them despite the inaccuracies. Hence, these algorithms can be integrated into citation matching algorithms. However, to know what kind of permutations the algorithms need to perform, a deeper understanding of the inaccuracies occurring and their characteristics is necessary.

2.2.4 Missed citation

A missed citation is one that could not be matched to its corresponding cited article and, therefore, is not considered in bibliometric calculations. Jacsó (2008d) further distinguishes between orphan and stray references. An orphan reference is one that has no master record in the respective database, i.e. the cited article is not indexed by it. A stray reference has a master record, but was not matched correctly to it, which is what we consider to be a missed citation

in this doctoral research. Missed citations are also sometimes referred to as lost citations (Moed, 2002). WoS provides a useful feature for identifying missed citations in its database: the *Cited Reference Search*. The feature allows searching for variations of *author name*, *publication name*, *publication year*, *volume*, *issue*, *pages* and *title* and provides a list of citations that match the variations found in stray or orphan references. Therefore, one can identify potential missed citations in the system, validate them manually and add them to one's citation analysis. Figure 4 shows an example of a missed citation in the cited reference information of an article in WoS, which holds an incorrect page number, as opposed to Figure 5 which shows a citation to the same cited article that was correctly matched. In contrast to the matched citation, the missed citation does not include the complete bibliographic information and is not linked to the respective WoS record.

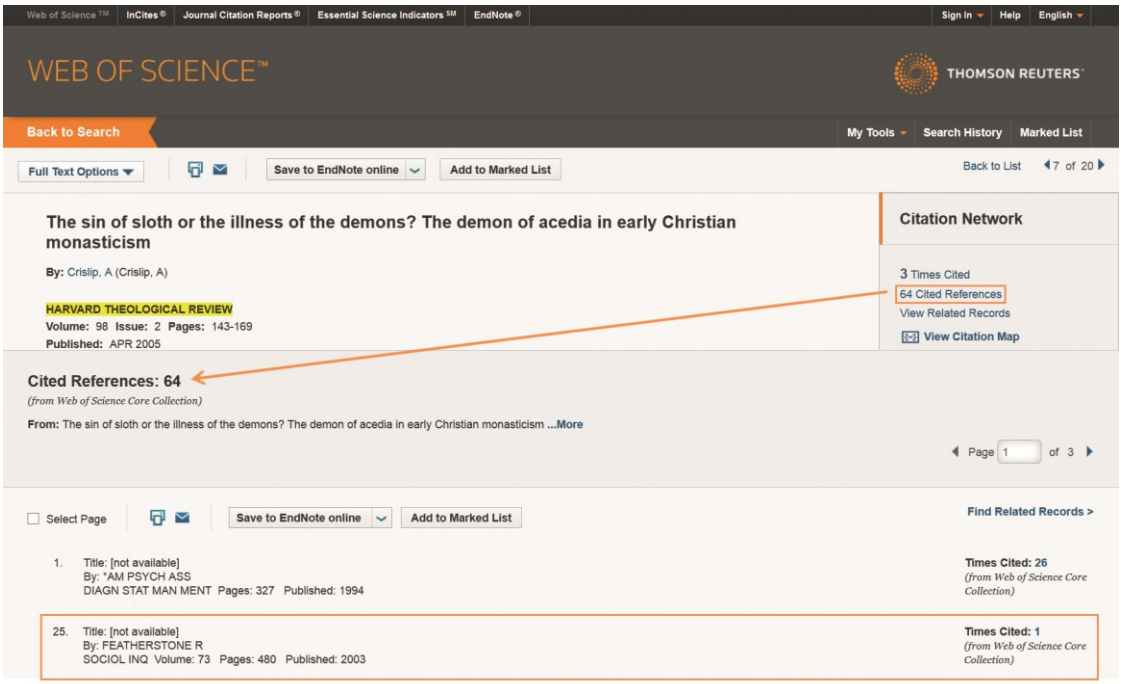


Figure 4: Example of a missed citation in the cited reference information of an article in WoS

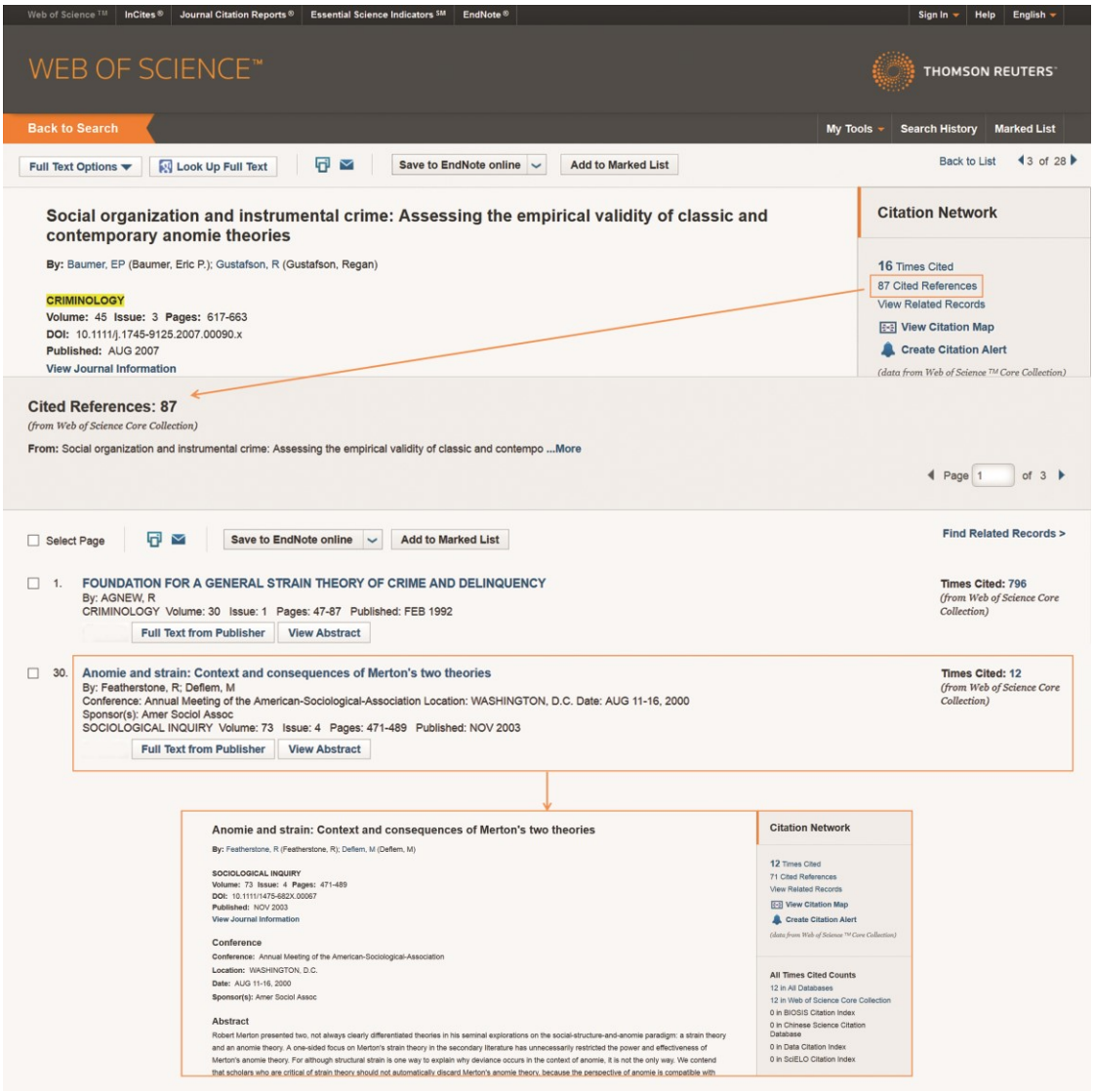


Figure 5: Example of a correctly matched citation in the cited reference information of an article in WoS

The reasons why some citations are not matched to their corresponding target articles can be author-induced errors in the references, e.g. errors in journal volume numbers or starting page numbers, or flaws in the data-handling or matching process, or both. Particularly problematic are references to publications written by consortia or by authors from non-English-speaking countries, research papers published in journals with dual volume-numbering systems or combined volumes, as well as journals applying different article numbering systems (van Raan, 2005). The reasons are further discussed in section 4.2.

2.3 Bibliometric indicators

Bibliometric indicators are the results of bibliometric studies. The objects of investigation are commonly:

- quantity indicators: the number of publications that indicate research output (per institution, field, researcher, etc.)
- impact indicators: the number of citations that these publications have received to measure scientific impact or performance of the research output
- structural indicators: co-authorship to measure the extent of (international) collaboration and intellectual linkages between researchers, institutions, countries, etc.

For these three categories, different indicators can be calculated. They can be as simple as the quantitative indicators for the number of published papers or the number of cited papers. However, the most commonly used ones are the performance (or impact) indicators Journal Impact Factor (JIF) and the *h*-index. The JIF provides the average citation rate for one- and two-year-old articles published in a particular journal and was invented by Garfield (1972) to measure the frequency with which the average article in a journal has been cited. The *h*-index is a bibliometric indicator that measures an individual's scientific research output. It "gives an estimate of the importance, significance, and broad impact of a scientist's cumulative research contributions" (Hirsch, 2005, p. 16572). A researcher with an index of *h* has published *h* papers, each of which has been cited at least *h* times and, therefore, provides a balance between productivity and citedness. Structural indicators usually calculate co-citation maps that indicate collaboration.

Research on bibliometric indicators is ongoing and indicators, such as the *g*-index (Egghe, 2006), the Eigenfactor (Bergstrom, 2007), the crown indicator (Leiden Ranking in 2007), the new crown indicator (Waltman, van Eck, van Leeuwen, Visser & van Raan, 2011), etc., are newly invented and critically investigated (Costas & Bordons, 2008; Davis, 2008; Franceschet, 2010b; Lundberg, 2007). We do not explain every bibliometric indicator in detail, as this goes beyond the scope of this dissertation. In the context of this doctoral research, data accuracy is a dominant factor for all types of bibliometric indicators, as accurate data ensure the correct matching of articles and consequently the correct calculation of indicators. Hence, data accuracy is even more important for indicators that employ citation analysis as they rely on the correct matching of citing articles to their cited articles. However, some researchers argue that relative bibliometric indicators, such as the *h*-index, should be robust enough to provide

accurate results even though not all citations might be considered (Jacsó, 2009; Franceschini & Maisano, 2011; Henzinger, Suñol & Weber, 2010). Yet again, this also depends on the level of granularity of the study as well as the data sources used (Henzinger et al., 2010). A researcher's *h*-index could be more influenced by missing citations than the *h*-index of an entire research unit, university or country.

2.4 Bibliometric data sources

Bibliometric data sources are the sources of bibliographic data used in citation analyses. In this research, we distinguish between three kinds of bibliometric data sources: 1) bibliographic references, which are the root of all citation analyses, 2) citation indexes, which process publications and their references to provide basic bibliometric indicators and raw citation data, and 3) applied bibliometric research groups which build on these citation indexes and apply their own in-house methodologies to match the data provided.

2.4.1 Bibliographic references

Since part of the scientific communication process of publishing one's research is citing other researchers' work and ideas, citations are a form of acknowledgement whereby the ideas are either further evolved or sometimes refuted (Bornmann & Marx, 2013). The references to other researchers' publications are documented in the bibliographies of one's own scientific publications. Hence, reference lists are the raw material for carrying out citation analyses (Garfield, 1972; MacRoberts & MacRoberts, 1989; Dinkel, 2011) and can be defined as the first and most important bibliometric data source.

2.4.2 Citation indexes

With the establishment of the Institute of Scientific Information (ISI) in 1960 and the start of collecting scientific publications in a bibliographic database, Eugene Garfield laid the cornerstone of citation analysis. In this database not only the publication data, but also the citing references were indexed. However, initially the database was built as a literature retrieval database for journal articles (Hood & Wilson, 2003; Neuhaus & Daniel, 2008). The use as a source for citation analyses was a subsequent development when, a few years later, Garfield turned the index of references into an opportunity for tracking citations, and thus the SCI was born. Today, this database is known as WoS and it has found potential competitors in Scopus and GS. Besides these three main data sources, other domain-specific citation indexes

exist and have been used in comparative bibliometric studies complementary to WoS, Scopus and GS, e.g. Chemical Abstracts for chemical literature (e.g. Whitley, 2002; Neuhaus & Daniel, 2008), PubMed for medical literature (e.g. Falagas, Pitsouni, Malietzis & Pappas, 2008), PsycINFO for literature in the behavioral sciences and mental health (e.g. Bauer & Bakkalbasi, 2005; Jacsó, 2008a), CSA Illumina for SSH literature (e.g. Norris & Oppenheim, 2007) and CiteSeer for literature related to computer and information science (e.g. Bar-Ilan, 2006). However, in this research, we focus on the three main citation indexes, WoS, Scopus and GS. Their characteristics are discussed in this section.

WoS is the web portal provided by Thomson Reuters for searching three different citation indexes (Science Citation Index Expanded (SCIE), Social Sciences Citation Index (SSCI), Arts & Humanities Citation Index (AHCI)). As of January 2014, the former WoS, consisting of these three citation indexes, has been renamed Web of Science Core Collection. We continue to use the commonly known abbreviation WoS to refer to this Core Collection. The SCIE and the SSCI both cover publications as well as citations from 1900 to the present, whereas the AHCI covers publications back to 1975 and citations back to 1945 (Thomson Reuters, 2014b). The counterpart to WoS is Elsevier's Scopus, launched in 2004 as a reaction to the monopoly held by Thomson Reuters. Scopus covers bibliographic records and abstracts back to 1966 and citations back to 1996. In March 2014, Elsevier announced the launch of a project that will add citing references (back to 1970) to pre-1996 content (Elsevier, 2014a). Both databases offer functionalities for searching, browsing, sorting, saving and exporting records to citation management software, as well as citation counts and basic citation analyses. They are both subscription-based services. A cost-free alternative is GS, also launched in 2004. Contrary to WoS and Scopus, Google does not provide clear information about the number of records, indexed titles, document types, subject areas covered or the time span in its database, which makes comparability and quality control even harder than with the two commercial ones. Additionally, bibliographic records can only be downloaded manually.

Coverage. WoS and Scopus both cover a large variety of journals (Scopus: over 21,000 (Elsevier, 2014c); WoS: over 12,000 (Thomson Reuters, 2014a)), as well as an ever increasing number of books and conference proceedings. However, WoS and Scopus do not always provide constant coverage of indexed journals over time (Meho & Yang, 2007; Jacsó, 2008c) and sometimes articles and even entire issues of indexed journals are missing (Meho & Rogers, 2008; Vieira & Gomes, 2009; cf. section 5.5). Some studies also criticize that their coverage is still not large enough because they do not cover all scholarly literature (Harzing, 2008; Meho & Yang, 2007). However, on account of mathematical laws, such as the laws of Lotka, Zipf

and Bradford[10], that have been studied in the context of bibliometrics (e.g. Naranan, 1970; Rousseau, 1998; 2002; Egghe, 2005), Garfield (1972) argues that complete coverage is not necessarily important to determine scientific impact. Furthermore, WoS is claimed to be biased towards English-language publications and natural sciences (NS) (Kostoff, 2002; Meho & Yang, 2007; Harzing, 2008). Other studies corroborate this by stating that Scopus and GS provide better coverage of non-English-language publications (López-Illescas, de Moya-Anegón & Moed, 2008; Kousha & Thelwall, 2008) and also cover more social science literature than WoS (Norris & Oppenheim, 2007; Harzing, 2013a). On the one hand, GS is praised for covering a larger and more diverse amount of literature, such as more conference proceedings and other modes of scholarly communication like preprints from arXiv as well as publications from government and academic websites (Belew, 2005; Bauer & Bakkalbasi, 2005; Bakkalbasi, Bauer, Glover & Wang, 2006; Bar-Ilan, 2010). On the other hand, it is criticized for also covering non-scholarly literature, such as presentations, master theses, etc., which inflate citation counts (Jacsó, 2006; Levine-Clark & Gil, 2009; Harzing, 2008). However, some authors argue that the inclusion of non-scholarly citations as well as another limitation of GS, namely a large number of duplicates, can be attenuated by robust citation metrics, such as the *h*-index (Harzing & van der Wal, 2009; Meho &Yang, 2007; Vaughan & Shaw, 2008).

Scientific subject category. Every database has its own scientific subject category system. Therefore, the classification of journals according to WoS is not the same as that of Scopus. While the WoS classification is based on information extracted from journal titles (Moed, 1996), Scopus's way of classifying journal titles is not documented. GS, on the other hand, does not provide any subject classification for its publications at all.

Document type. It depends on the field of the bibliometric study to decide which document types should be included in an analysis. Yet, one needs to be aware that the classification of document types can differ between data sources, such as WoS and Scopus, and that different disciplines may employ different interpretations of document types (Archambault et al., 2009). For instance, in Physics and Astronomy, letters can report truly original research findings, whereas in other disciplines letters in journals are rather a means to comment on another person's work (Moed, 1996; Moed & van Leeuwen, 1995). In Scopus, the classification of documents is not really clear and Harzing (2013b) reports on several attempts to obtain

[10] These mathematical laws describe statistical effects which can be applied to bibliometric studies and prove, for example, that only a small number of researchers publish the majority of publications. For a detailed discussion of these laws and their relation to bibliometric studies, cf. for example Havemann (2009).

additional information from Elsevier which remained unanswered. WoS automatically classifies any research article with more than 100 references as a review, which can cause problems in social science disciplines where it is common to have original research articles with more than 100 references and review articles are not acknowledged as original research (Harzing, 2013b). Another interesting misclassification was observed in articles that included a note like "part of this paper was presented at a conference" or even "this is based on a paper previously presented at a conference", which were then classified as proceedings papers (Harzing, 2013b). In the meantime, WoS has canceled this rule.

Publication year. Before WoS had an online version, the CD-ROM versions left room for interpretation of publication years. In that version, each record also had a database publication year assigned, which marked the year when the document was added to the database (Jacsó, 1995 & 1997). In the online version, this discrepancy is no longer an issue. Yet, it is still important to clearly define the publication years of the cited articles considered and not to confuse them with the citation period, i.e. citation window, which corresponds to the publication years of the citing articles (cf. section 2.2.2).

Comparison of databases. In an effort to determine which of the three main bibliometric data sources (WoS, Scopus and GS) is the best fit for bibliometric analyses, studies have compared these data sources with regard to coverage and overlap of publications and citation counts. In terms of coverage, the majority of such works have juxtaposed the available formats, i.e. publication and document types, temporal, i.e. publication years or citation windows, and geospatial coverage, i.e. investigation of country-specific journals or languages, as well as the extent to which the domains (NS vs. SSH) or specific disciplines are covered. Building on these facets of coverage, studies have compared the overlap of publications and the corresponding citation counts (e.g. Bauer & Bakkalbasi 2005; Cameron, 2005; Meho & Yang, 2007; Mingers & Lipitakis, 2010; Adriaanse & Rensleigh, 2013). Others have investigated and compared rankings of research units according to bibliometric indicators based on citation data from different data sources (e.g. Bar-Ilan, 2008; López-Illescas et al., 2008; Meho & Rogers, 2008; Sanderson, 2008; Jacsó, 2009; Franceschet, 2010a). Most of the studies report that the overlap of citations is higher in the case of WoS and Scopus (between 58 to 70%) than when these two data sources are compared with GS citations (Meho & Yang, 2007; Bar-Ilan, 2010; Jaćimović, Petrović & Živković, 2010). The same is true for the results of bibliometric calculations and rankings. Authors report highly correlated results for WoS and Scopus and slightly different results for GS (Bauer & Bakkalbasi, 2005; Bar-Ilan, 2008; López-Illescas et al., 2008; Archambault et al., 2009; Franceschet, 2010a).

In general, the majority concludes that WoS and Scopus should be used complementarily in bibliometric studies (Sanderson, 2008; Li, Burnham, Lemley & Britton, 2010) and that the choice of database depends on the purpose of the study, the research field in question, the types of documents to be investigated, the types of journals to be included (e.g. peer review, open-access journals) and on whether pre-1996 citations are required (e.g. Frandsen & Nicolaisen, 2008; Neuhaus & Daniel, 2008; Meho & Rogers, 2008; Bar-Ilan, 2008; Bar-Ilan, 2010; Mingers & Lipitakis, 2010; Adriaanse & Rensleigh, 2013). Moreover, scientific areas, such as mathematics, engineering, economics and social sciences, arts and humanities, where journals play a less central role as a scholarly communication system (Moed, 2005), will require different data sources than WoS and Scopus. Norris & Oppenheim (2007) suggest that, for studies of social science literature, WoS should be replaced by Scopus, while Kousha & Thelwall (2007) are a bit more cautious and recommend GS for studying citations in the social sciences, but they also admit to having found some exceptions and hint that replacing the traditional data source WoS by GS citations would be problematic.

2.4.3 Applied bibliometric research groups

The third kind of bibliometric data source is applied bibliometric research groups that perform bibliometric analyses mainly on behalf of organizations with a stake in science and technology, such as national governments or national and international funding agencies. Three of the most renowned applied bibliometric research groups are: Centre for Science and Technology Studies in Leiden (CWTS), Institut für Forschungsqualität in Berlin (iFQ) and Science-Metrix (SM) in Montreal. They provide bibliometric services and products closely related to research evaluation, which are often based on citation analyses[11]. CWTS, iFQ and Science-Metrix use raw WoS data through a bibliometric production platform licensed by Thomson Reuters. Depending on the scope of the study, Science-Metrix additionally employs Scopus data, which also allows them to use the article title in the citation matching process (P. Deschamps, personal communication, March 4, 2014). CWTS and Science-Metrix match the citation data according to the matching algorithms they have developed. As mentioned in section 2.2.3, iFQ's matching algorithm is still under development. For the time being,

[11] e.g. http://www.science-metrix.com/pdf/SM_INAC_Bibliometrics_Arctic_Research.pdf or http://www.leidenranking.com/

therefore, they rely on the data as matched by WoS, complemented by citations found through the *Cited Reference Search*[12].

2.5 Summary

In this chapter we have defined the fundamental bibliometric terminology used in this doctoral research. Citation analysis is the most important informetric method in research evaluation and, therefore, the focus of this dissertation. In the context of citation analysis, we explained the importance of accurate linkage between target and source articles, which is accomplished in citation matching processes. Citation matching usually employs a set of different match keys and may also make use of string matching methodologies that can cope with possible discrepancies in the references. A missed citation is a citation that could not be correctly matched to its cited articles in this process. WoS provides a feature called *Cited Reference Search* that allows searching for such missed citations in the system. Furthermore, we gave a short overview of the commonly used bibliometric indicators and underlined their dependence on data accuracy. Last, we defined the concept of a bibliometric data source not only as a classic citation index, such as WoS, Scopus and GS, but widened the notion to include the underlying raw material, namely bibliographic references, as well as the bibliometric services of applied bibliometric research groups.

[12] Appendix A lists information, acquired in personal communications, about the citation matching algorithms insofar as the applied research groups allowed publication of this information for reasons of competitive advantages.

3 Defining Data Accuracy

In order to define what data accuracy means for a bibliometric data source, we continue the theoretical part of this doctoral research by exploring the existing definitions of data accuracy, its generic concept, data quality, as well as its antonym, data inaccuracy. In sections 3.1 and 3.2, we summarize definitions of these three concepts and formulate our own definition of what data inaccuracy comprises in this research in order to be able to assess data accuracy. In sections 3.3 and 3.4 approaches to assess (bibliographic) data accuracy are discussed. Section 3.5 summarizes the chapter.

3.1 Data quality

As data accuracy is one specific aspect of data quality, this section commences by shedding light on the term *data quality*. The origin of the word data is the Latin noun *datum*, meaning *something given*. The Oxford English Dictionary (2013) defines data as "an item of information" or "information in digital form". The definition of quality in the ISO 9000 standard is "the totality of features and characteristics of a product, process or service that bears on its ability to satisfy stated or implicit needs" (ISO 2005). In other words, data quality can be defined as "fitness for the purpose of use" (Wang & Strong, 1996, p. 6; Maydanchik, 2007, p. 245) of an item of information.

In the context of a database, Data Quality (DQ) is a very complex concept to describe and especially to measure. Redman (1996) developed a system-centered framework that defines the dimensions of data quality according to three aspects of data: data modeling, data values and data representation. He focuses on data per se and disregards other aspects of data quality, such as storage and security. His framework can be applied to a variety of databases as it deals with errors that can be measured formally and it provides the basis for the widely accepted and used categorization of data value quality into the four dimensions *accuracy*, *currency* (sometimes also referred to as *timeliness*), *completeness* and *consistency* (e.g. Wand & Wang,

1996; Wang & Strong, 1996; Naumann, 2002; Jarke, Lenzerini, Vassiliou & Vassiliadis, 2003; Bovee, Srivastava & Mak, 2003; Batini & Scannapieco, 2006). Although not all researchers agree on the exact same definitions of quality dimensions of data values (*accuracy*, *currency*, *completeness* and *consistency*), the essence of each is the same. The following paragraphs give short explanations of the four dimensions, mainly based on Redman (1996) and complemented by the above-mentioned literature.

> ***Accuracy.*** Accuracy refers to whether data values are correct or not. It is not easy to quantify data accuracy, as a standard or correct value to compare data with may not be available. The suggested formula to calculate data accuracy (p) is to divide the number of correct values by the number of total values: p = the number of correct values / number of total values.
>
> ***Currency.*** Data values can change over time. Currency refers to the degree to which data is up-to-date. This means that data currency is a special case of data accuracy. The concept of currency is, therefore, only applicable to changing entities in the database, i.e. for bibliometric data sources this is, for example, the citation count in the field *Times cited*.
>
> ***Completeness.*** Attributes in a database can be mandatory, optional or inapplicable. Therefore, *null* in an attribute can have different meanings, which needs to be considered when assessing the completeness of data values.
>
> ***Consistency.*** Overlapping data need to have consistent values. For instance, the name of an institution must have the same string in every record. Furthermore, data values also need to be consistent with other values: the name of an institution must match its country and the city must match the country and the zip code.

In most databases, trade-offs between data dimensions have to be made (Batini & Scannapieco, 2006). For example, if one decides in favor of accurate (or complete or consistent) data this may adversely affect currency, as it takes time to check the accuracy of data. Web applications often opt for current data and as a consequence neglect the other three dimensions. The choices can differ in different domains and business contexts. However, most studies have identified accuracy of data values as the key dimension of data quality (Wand & Wang, 1996; Batini & Scannapieco, 2006) and as "one of the main intrinsic properties of data" (Naumann, 2002, p. 30; Bovee et al., 2003; Wang & Strong, 1996). In bibliometric databases (cf. section 2.4) and citation analysis (cf. section 2.2), all four dimensions of data value quality are important: on the one hand, accurate and consistent data values ensure a correct citation matching process (cf. section 2.2.3), which in turn contributes to a complete and up-to-date

(currency) publications and citations count. Accurate data values of bibliographic references are especially important for the syntactic matching of citations and are, therefore, the basis of a successful citation matching process. Consistent values play a role in the semantic matching process, e.g. author name disambiguation (cf. section 4.1), and are undoubtedly essential to the process as well. In this research, we decided to focus on the dimension that first and foremost impacts the citation matching process: the accuracy of bibliographic data values.

3.2 Data accuracy – data inaccuracy

In the literature, the terms error rate, correctness, reliability, integrity and precision are often used as synonyms for data accuracy (Naumann, 2002). Data accuracy is defined as

> "…the recorded value is in conformity with the actual value" (Ballou & Pazer, 1985, p. 153).
>
> "[…] the nearness of the value *v* to some value *v'* in the attribute domain, which is considered as the correct one […]" (Redman, 1996, p. 255).
>
> "[…] the extent to which data values are in conformance with the actual or true values" (Wang & Strong, 1996, p. 18).
>
> "[…] the avoidance of errors in all stages of creating an information unit: (a) in document analysis; (b) during entry in the data fields; and (c) orthographical errors" (Rittberger & Rittberger, 1997, p. 27)
>
> "[…] the extent to which data is correct and reliable" (Kahn, Strong & Wang, 2002, p. 187).
>
> "[…] the quotient of the number of correct values in a source and the overall number of values in the source. A value is an instance of an attribute" (Naumann, 2002, p. 30).
>
> "[…] information being true or error free with respect to some known, designated, or measured value" (Bovee et al., 2003, p. 59).
>
> "[…] the closeness of the value in our database to the true value" (Dasu & Johnson, 2003, p. 105).
>
> "[…] the validity of the data with respect to the real-world values" (Jarke et al., 2003, p. 155).
>
> "[…] the closeness between a value *v* and a value *v'*, considered as the correct representation of the real-life phenomenon that *v'* aims to represent" (Batini & Scannapieco, 2006, p. 20).

> "[…] the correctness of data items, compared to a baseline" (Even & Shankaranarayanan, 2007, p. 83).

What these definitions have in common is that they define data accuracy as the extent to which values are correct and the correctness of the values should ideally correspond to the real-world values. For example, Table 1 represents correct, i.e. real-world, values of a dissertation database. This means that, in the real world, i.e. in the paper or electronic versions of these dissertations, the titles of the dissertations and the author names correspond to the values in this database. Batini & Scannapieco (2006) further distinguish between syntactic and semantic data accuracy. According to their definition, DQ methodologies usually consider syntactic accuracy and define it as the closeness of a value *v* to the elements of the corresponding definition in domain *D*. Syntactic accuracy is not necessarily interested in comparing *v* with its real-world value *v'*, but checks whether *v* is any of the values in *D*, or how close it is to values in *D*. On the other hand, semantic accuracy relates to the concept of correctness (Batini & Scannapieco, 2006).

Table 1: Example of a dissertation database

ID	Title of dissertation	Author name
1	Country and Language Level Differences in Multilingual Digital Libraries.	Maria Gäde
2	Data Accuracy in Bibliometric Data Sources and its Impact on Citation Matching.	Marlies Olensky
3	From Curation to Collaboration. A Framework for Interactions in Cultural Heritage Information Systems.	Juliane Stiller

If the author names in tuples 1 and 2 of our dissertation database (Table 1) were switched, a semantic error would occur. Yet, the author names would still be syntactically correct as both author names are admissible in the domain of authors of dissertations. A syntactic error would occur if, for example, the author name in tuple 1 was spelled *Maria Gede* instead of *Maria Gäde*. Hence, syntactic accuracy is measured by distance functions; semantic accuracy should be measured by domains like <yes, no> or <correct, incorrect> (Batini & Scannapieco (2006).

Few authors have formulated individual definitions of what actually constitutes data **in**accuracy, even though it is necessary to know what qualifies data as accurate or inaccurate in order to measure data accuracy. Wand & Wang (1996, p. 93) define inaccuracy as "a result of a garbled mapping into a wrong state of the information system". Batini & Pernici (2006, p. 52) adopted their definition and added "[…] where it is possible to infer a valid state of the real world though not the correct one". Jacsó (1995, p. 150) describes inaccuracy as "a

euphemism for erroneous, wrong data". Moed, one of the few researchers who have actually studied inaccuracies in bibliometric data sources, uses the term discrepancy "to indicate […] differences or variations between a target article intentionally cited in a reference and the citing reference itself" (Moed, 2005, pp. 173-174).

We do not agree with Jacsó's definition because, as Moed points out, a difference between two bibliographic data records does not necessarily need to be an error, but could be due to technical coding differences or different transliterations, punctuation, etc. Hence, we inferred a definition for data inaccuracy for this research from the definitions of data accuracy and inaccuracy in the literature (mainly based on Redman, 1996; Moed, 2005; Batini & Scannapieco, 2006):

> *Data inaccuracy describes a discrepancy between the correct value v' and the assessed value v, i.e. any non-conformity between these two values is recorded. The term data inaccuracy is used synonymously with discrepancy. A data inaccuracy is not necessarily an error.*

3.3 Data accuracy assessment

The literature provides a variety of techniques to assess DQ in databases and summarizes them in different DQ assessment frameworks (e.g. Batini, Cabitza, Cappiello & Francalanci, 2008; Even & Shankaranarayanan, 2007; Lee, Strong, Kahn & Wang, 2002; Su & Jin, 2004; Scannapieco, Virgillito, Marchetti, Mecella & Baldoni, 2004; Wang, 1998). These frameworks mostly describe how enterprises can maintain quality in their databases by employing record linkage, business process rules and similarity measures (Batini, Cappiello, Francalanci & Maurino, 2009).

The measurement of data accuracy in these frameworks is basically defined as the ratio of correct to incorrect values and can be expressed in different ways (cf. Table 2). The definition of what a data unit consists of depends on the individual assessment process. It could be a data field, data record or even an entire dataset. Rittberger & Rittberger (1997, pp. 33-34) have taken a slightly different approach and suggested measuring the error rate in bibliographic online database production as the "number of errors per 1,000 entered symbols or number of errors in a specific data field". However, before one can express data accuracy as a metric, one

needs to define what the correct value is. Therefore, the more important questions are: what qualifies as an incorrect data value? How can incorrect values be identified?

Table 2: Data accuracy measurement

Definition	Formula	Source
Accuracy	$\frac{incorrect\ values}{correct\ values}$	Loshin (2001)
Free-of-error dimension	$1 - \frac{count\ of\ data\ units\ in\ error}{total\ number\ of\ data\ units}$	Pipino, Lee & Wang (2002)
Free-of-error rating	$1 - \frac{number\ of\ data\ units\ in\ error}{total\ number\ of\ data\ units}$	Lee, Pipino, Funk & Wang (2006)
Accuracy score	$\frac{count\ of\ rel.rec. - count\ of\ err.\ rec.}{count\ of\ relevant\ records}$	Maydanchik (2007)
Syntactic accuracy	$\frac{number\ of\ correct\ values}{number\ of\ total\ values}$	Batini et al. (2009)

As the answer to the first question can vary for each database depending on its content, the literature cannot provide any universal answers. The correct values for data accuracy assessment must, therefore, be determined case by case[13]. Assuming we have defined the correct values, against which the data will be assessed, and follow the definition of data inaccuracy from the previous section 3.2, incorrect or discrepant data values can be identified by distance or similarity functions (e.g. Redman, 1996; Batini & Scannapieco; 2006). For example, the Levenshtein distance function is a widely used method to measure the distance between two strings, i.e. it measures the number of edits the function has to perform to transform a string *s* into string *s'* (Levenshtein, 1966). In contrast, the Jaro-Winkler string comparator (Winkler, 1995) measures the similarity of two strings, i.e. how many characters two strings have in common. Other, more sophisticated edit-distance functions or algorithms used in fuzzy string matching methodologies were listed in section 2.2.3 on citation matching, since they go beyond a mere assessment of data values but apply certain permutations of data values to match them despite possible inaccuracies.

3.4 Bibliographic data accuracy assessment

This doctoral thesis investigates the accuracy of references, i.e. bibliographic data. Therefore, we looked for accuracy assessment methodologies in the literature that have been applied to

[13] For this doctoral research, they are defined in section 5.1.

bibliographic data values. The accuracy of references in research articles and databases has been studied before, but not by employing any of the above-mentioned frameworks from the data quality literature. Therefore, we compiled a complete list of these studies (98 studies in total) and analyzed them to ascertain whether there is a standardized and/or automated way to assess and categorize inaccuracies in references (Olensky, 2012). The main aspects of evaluation were: main goal of study; subject area; data sources; number of journals investigated; number, publication type and year of citing articles; number and publication type of cited articles; selection of the data sample; assessment method; error categories. The studies were mainly conducted by researchers in their own field to call attention to inaccuracies and negligent references that would impede fellow researchers from retracing their research process and sources. Furthermore, a few studies (e.g. Moed & Vriens, 1989; Moed, 2005; García-Pérez, 2010) in this evaluation assessed data accuracy of citation indexes.

The results revealed that, in most cases, bibliographic data is measured by the accuracy of the following fields: *author name(s)*, *journal title*, *volume*, *year* and *pagination* (Table 3). The majority of the studies used the original publication as the *gold standard* for verification, i.e. as the correct (real-world) values *v'*. Even the database studies consulted the original articles in most cases, except for two of the studies which employed match keys, as used in citation matching processes, to identify inaccurate records. If we not only wish to identify inaccuracies, but also need to determine whether the mistakes were made by the author or introduced by the database, it is necessary to check the references from the original citing articles against the existing data. Therefore, it depends on the intended aim of the assessment whether the original needs to be consulted or not.

The study also showed there is no standardized way of categorizing bibliographic data errors and the granularity of categories varies. Half of the studies divided the errors into the groups *major* and *minor*; some added an *intermediate* category (Olensky, 2012). However, the studies do not fully agree on what qualifies as a major, intermediate or minor error. Other studies, including the studies investigating the accuracy in citation indexes, listed the errors describing in which field they occurred and partially describing the nature of the discrepancy (e.g. page number missing, small variation in author name (Moed, 2005); wrong cited year, swapping of digits (Larsen et al., 2007)).

Table 3: Aspects of bibliographic data accuracy (Olensky, 2012)

Bibliographic field	% of studies
author name(s)	100%
author initials	76%
author number	54%
author order	39%
article titles	97%
journal title	100%
volume	100%
issue	17%
year	98%
pagination	100%

Apart from the above-mentioned match-key studies, bibliographic data accuracy has not been assessed in an automated way before (Olensky, 2012). Since match keys can identify inaccurate records, but do not provide information about how inaccurate data is, we tested the Levenshtein distance function (LDF) as an accuracy assessment method for bibliographic data in a pilot study (Olensky, 2013). The LDF measures the distance between two data values and indicates the number of edits to transform one value into the other (in contrast to the aforementioned Jaro-Winkler string comparator), which correlates with the widely used definitions of accuracy scores in the DQ literature. The study investigated whether the automated assessment method, as described in the DQ literature, can be applied to bibliographic data. The main result is that the Levenshtein distance function is a good means to determine whether a data record contains discrepancies, but the score does not provide a true picture of how inaccurate a bibliographic data value is unless additional rules are applied. For example, the LDF produces high scores for article titles whenever the subtitle is missing in the reference or when the titles are translations of each other, yet, this does not necessarily indicate a major inaccuracy. To counterbalance certain shortcomings of the LDF, we evolved a set of rules during a manual assessment process that takes into account characteristics of bibliographic data and their sources. The rules spelled out in the manual assessment method reflect most of the required adjustments to be made to an automatic assessment method. They mirror specific characteristics of bibliographic data (Olensky, 2013):

- different presentation of data (e.g. one-page articles in Scopus have no ending page)
- abbreviated publication names
- translated article titles

- punctuation
- special characters (e.g. German Umlaut)
- non-alphanumeric characters (e.g. α)
- domain-specific abbreviations (e.g. *Ag* // Silver)
- different weighting of inaccuracies (omitted // inaccurate // incomplete)

Even though the findings of this study are a first step towards an automated accuracy assessment of bibliographic data, the data sample investigated was too small to present a comprehensive list of data manipulation rules that need to be considered in an automated assessment process. Thus, we identified the need for an in-depth analysis of bibliographic data, its characteristics and the inaccuracies occurring therein.

3.5 Summary

The literature provides different definitions of data accuracy metrics, but their essence is the same: once a way has been found to identify inaccurate data values, the data accuracy of a database can be defined as the ratio of the inaccurate data values to the accurate data values. Optionally, to measure how (in)accurate values are, a distance or similarity function, such as the Levenshtein or Jaro-Winkler function, can be used. The bibliographic data accuracy of references in research articles is assessed by consulting the original article the reference cites and investigating the bibliographic fields *author name(s)*, *journal title*, *volume*, *year* and *pagination*. Hence, these findings as well as our definition of what data inaccuracy comprises influence the methodological considerations of this doctoral research.

4 Inaccuracies in Bibliometric Data Sources

Chapter 2 explained the concept of a bibliometric data source as understood in this research, and data accuracy and inaccuracy as well as cases of bibliographic data accuracy assessment were discussed in chapter 3. This chapter explains how the accuracy of a bibliometric data source can be understood and discusses the current state of research on inaccuracies of bibliographic data values. Section 4.1 draws a general picture of what influences accuracy or, more specifically, data accuracy in a bibliometric data source and justifies the focus of our data analysis described in chapter 5. Section 4.2 discusses inaccuracies in bibliographic data fields which have been the subject of previous studies and play a primary role in the citation matching process. The different sources of inaccuracies as well as the predominant types of inaccuracies are discussed. Section 4.3 concludes this chapter.

4.1 (Data) accuracy in bibliometric data sources

In personal communication with fellow researchers and established bibliometricians (e.g. Frank Havemann, Paul Wouters, Stefan Hornbostel), it becomes apparent that the data accuracy of bibliometric data sources is first and foremost understood as the correctness of the citation counts in databases like WoS or Scopus. However, citation counts are the result of bibliographic data values extracted from references and of the process applied to match these values to the respective target articles. This section explains the existing relations between the three bibliometric data sources defined in section 2.4 and explains why the focus of this dissertation targets the data accuracy of bibliographic references.

As the producer and user of a bibliometric study, one must be able to answer a few crucial questions about the data used in the calculation of citation counts, such as how the publication data was collected and how citations were matched to their target articles (Moed, 2002). However, aside from the few match-key studies discussed in sections 2.2.3 and 4.2, not one

comprehensive study provides information on the accuracy of citation matching in the different citation indexes available. Hence, the citation matching process is an immanent characteristic of the database and does not usually influence the choice of which citation index is used in a bibliometric study. The case is, of course, different for the applied bibliometric research groups, which match the data according to their algorithms developed in-house. Nevertheless, the user of a bibliometric statistic compiled by one of the research groups does not receive any information about the citation matching process applied either, but simply has to rely on its accuracy.

The citation matching process is in turn impacted by the accuracy of data values in the database, i.e. bibliographic data of the articles and citations (Moed, 2002; 2005). Figure 6 gives an overview of how the three bibliometric data sources, as defined in section 2.4, are related to and influence each other. Starting at the bottom of the figure, the bibliographic data in a publication consists of the main bibliographic data, such as article title, author name(s), institutional affiliation, address, and journal-specific data, e.g. journal title, volume number, and the data of the bibliographic references, which are then reflected in the cited reference information of the respective database. In the data ingestion process these values are extracted and assumably subjected to data cleaning and harmonization processes (cf. section 4.2). While parts of the main bibliographic data, such as author name(s) and journal title, are used in the citation matching process, the article title serves retrieval purposes only. Institutional affiliations and addresses in articles can be used to select a specific set of publications in bibliometric studies. For example, an analysis of the collaboration of two research institutions (or countries or authors) may use the institutional affiliations and addresses to disambiguate and uniquely attribute publications in the data sample used. The data accuracy of a bibliometric data source, therefore, ultimately hinges on the accuracy of the bibliographic references in a publication or in its database.

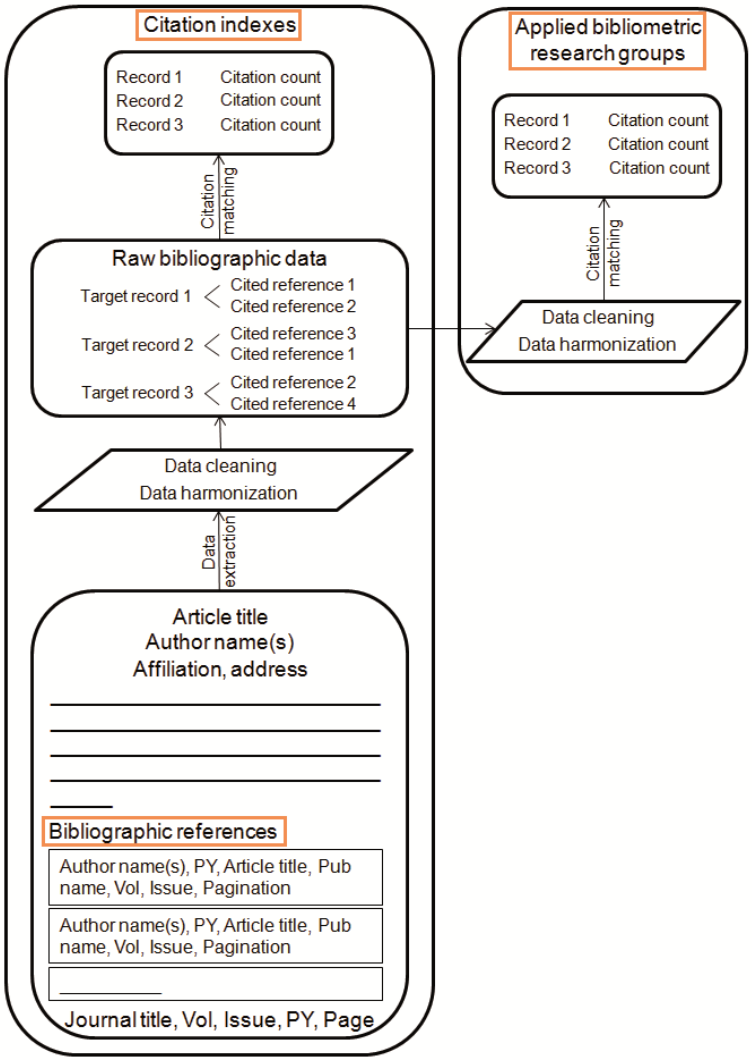


Figure 6: The relations of the three bibliometric data sources: bibliographic references, citation indexes, applied bibliometric research groups

In our analysis, the citation matching process is carried out by WoS, Scopus, GS and by the third kind of bibliometric data source, namely the applied bibliometric research groups. In WoS and Scopus the algorithms are reported to be more conservative than in GS (Larsen et al., 2007), but more sophisticated in the case of the applied bibliometric research groups (Neuhaus & Daniel, 2008). Section 2.2.3 explained that, in the citation matching process, typically the bibliographic data fields of first author (last name, first and second initial), publication name, publication year, volume number and starting page are employed. Hence, in section 4.2, we discuss inaccuracies related to the values in bibliographic data fields which have a primary impact on the citation matching process. Most of them are of a technical nature (e.g. typographical errors, spelling variations); others are more semantic in nature (e.g. incorrect interpretation of author names).

Semantic challenges in bibliometric data sources, that is in citation indexes and in the applied bibliometric research groups, are related to distinguishing publications, publication names, authors and their affiliations uniquely and identifying duplicate records. Another complex question concerns what defines a duplicate record. Do translations or pre-prints count as duplicates? How can they be identified? Furthermore, multiple manifestations, as defined in FRBR[14], of the same work or idea are quite a common phenomenon in computer science (Bar-Ilan, 2006). This does not refer to multiple entries in a database that fail to point to the same publication, but to ideas which are first published, for example, in conference proceedings and later in a journal. Sometimes, the publication may also contain slight changes, which makes it even more difficult to distinguish between different manifestations and expressions in the sense of FRBR. A work-around, which does not really solve this problem, is reported in a study by Meho & Rogers (2008) who declared that, if two works had the exact same title and were published within one year, they would treat the two publications as one.

Due to time and resource restrictions in the present research, we only address semantic challenges of citation analysis insofar as they concern the data values of the aforementioned bibliographic fields typically used in the citation matching process. Synonymic author names are discussed in section 4.2.1 with respect to incorrect interpretations of author names or incorrect first initials due to the use of a nickname instead of the correct first given name. However, the topic of author name disambiguation, which deals with homonymic author names as well as synonyms caused by changes in marital status or for religious or legal reasons (Bennett & Williams, 2006) or publications citing the name of a consortium rather than the actual author names (Moed, 2005), goes beyond the scope of this research. For discussions of homonymic author names cf. Aksnes (2008); for different solutions to disambiguate author names cf. the works of, inter alia, Companjen (2013) on probabilistic author name matching; Levin, Krawczyk, Bethard & Jurafsky (2012) on self-citation analysis; On, Lee, Kang & Mitra (2005), Huang, Ertekin & Giles (2006), Strotmann, Zhao & Bubela (2009) and Velden, Haque & Lagoze (2011) on co-author analysis and distance metrics; D'Angelo et al. (2011) on clustering approaches by institutional affiliations and WoS subject categories; Han, Giles, Zha, Li & Tsioutsiouliklis (2004) on supervised learning approaches by paper titles and publication venue titles.

[14] FRBR stands for Functional Requirements for Bibliographic Records and helps distinguish "products of intellectual or artistic endeavor (e.g., publications)" in "the work, a distinct intellectual or artistic creation; the expression, the intellectual or artistic realization of a work; the manifestation, the physical embodiment of an expression of a work; the item, a single exemplar of a manifestation." (IFLA, 1998)

4.2 Inaccuracies in bibliographic data values with a primary impact on the citation matching process

Errors, variations and inconsistencies in author name, journal title, publication year, volume and starting page are the most commonly reported inaccuracies in the references of individual papers (Moed & Vriens, 1989; Jacsó, 2005c; Galvez & de Moya-Anegón; 2006; Galvez & de Moya-Anegón, 2007; Neuhaus & Daniel, 2008; Adriaanse & Rensleigh, 2013; Chang, McAleer & Oxley, 2011). More specifically, inaccuracies can be related to publications written by consortia, i.e. large groups of authors, (Moed, 2002; van Raan, 2005), journals with dual volume-numbering systems or combined volumes, and journals applying different article numbering systems (Moed, 2002; van Raan, 2005; Tunger, Haustein, Ruppert, Luca & Unterhalt, 2010). Problems caused by a misunderstanding of foreign languages (Sweetland, 1989) or author names from non-English speaking countries (van Raan, 2005) as well as citing and cited authors with different linguistic backgrounds (Moed, 2005) can also be sources of inaccuracies in citing references. Different studies present differing results on which the most inaccurate bibliographic fields are: article title, author name and publication year (Meho & Rogers, 2008); volume and page number (Jacsó, 2004); page number, author names and year (Hildebrandt & Larsen, 2008); volume number, followed by a double error in volume number and starting page, and with the fewest records only having a wrong starting page (Liang, Zhong & Rousseau, 2014). Table 4 gives an overview of the identified problem areas of inaccuracies in references.

Inaccuracies in bibliographic data can be induced either by the author (e.g. provides inconsistent versions of his institutional affiliation), the citing author (e.g. jumbles the order of the cited author names) or by the database (e.g. interprets the issue number as the volume number) (Moed & Vriens, 1989; Buchanan, 2006; Hildebrandt & Larsen, 2008). All can result in a non-link between cited and citing article. Buchanan (2006) attributes errors in names and publication titles to being author-induced and lists transcription errors and cited articles omitted from the list of citing references as examples of database mapping errors. Inaccuracies introduced by the citing author in the citing references may or may not be corrected by copy editors of the journal publisher (Meyer, 2008). Any remaining inaccuracies will, therefore, find their way into the actual publication, be it the paper version or an online version. When the metadata of a new publication is indexed in a citation index, such as WoS, Scopus and GS, the delivery as well as the data extraction process may introduce additional inaccuracies, or

inaccuracies may be corrected in the data cleaning and harmonization process[15]. With the transition to online availability of most publications, publishers provide the metadata records electronically, directly extracted from the source document (Meyer, 2008; Moed, 2005), which should result in more accurate data. Even though, in recent years the share of electronic metadata records should have increased, not all of the data is recorded in that way. A considerable number of journal articles is still scanned by OCR software which extracts the bibliographic data. The scanning process as such is error-prone, but different citation styles can also cause inaccurate data values (Meyer, 2008). Together with articles from older publications, for which no electronic data was available at indexing time, the metadata from scanned documents still seems to represent the majority in citation indexes (Moed, 2005).

Additionally, inaccuracies can be passed on from one bibliography to the next (Simkin & Roychowdhury, 2003; Cameron, 2005; Wallin, 2005; Liang et al. 2014) because the authors do not even read and retrieve the original article, they read the article, but still copy the reference from another author's reference list, or they fail to retrieve the original article and still want to cite it and, therefore, copy the reference (Wallin, 2005). Even though Simkin & Roychowdhury (2003) applied a mathematical model to prove that authors copy references from each other, if two references contain the same discrepancy there was no empirical evidence that these references were copied (Moed, 2005). However, a recent study by Liang et al. (2014) found three routes that reference errors take in citing articles: "Route 1. Citing a paper and copying its reference; Route 2: Copying a reference from another paper but without citing this paper; Route 3: Copying references from an earlier paper published by the author himself (herself) without rechecking the accuracy of the reference".

The term *error rate* is inconsistently used in studies of errors in citation indexes. In some cases the error rate refers to errors found in the citing references, i.e. in the references of the original article or in the cited reference information of the citation index (e.g. Larsen et al., 2007; Moed, 2005), others to the actual bibliographic data record, i.e. the cited article in the citation index (e.g. De, Jones, Brazier, Jones & Fenton, 2001) and others to the number of references that were missed because of errors in the references (e.g. Chang et al., 2011).

[15] Our attempts to obtain information from Thomson Reuters, Elsevier and GS on their data ingestion processes remained unanswered. We can only quote E. Garfield in an interview with P. Jacsó (2004): the data cleaning process in WoS "is not a trivial one".

Table 4: Bibliographic inaccuracies (Garfield, 1981; Hood & Wilson, 2003; Moed, 2005; Meho & Yang, 2007; Harzing, 2008; Jacsó, 2008a, 2008b, 2008c, 2008d; Larsen et al., 2007; Tunger et al., 2010)

Area of concern	Problem description	Example from the data analysis / literature
Inconsistent and erroneous spelling of author names	Author names with special characters or diacritics	Suñol or Stalnionienė
	Double middle initials with or without punctuation	Weng, C-H vs. Weng, C.H. vs. Weng, CH
	Names with adjacent consonants or ligatures because of OCR errors	Gornis vs. Gomis
	Compounded names (with prefix, suffix or two or more parts)	van Hooland, S Padma Malar, EJ Zhang, Hongbao is indexed as Zhang, HB, instead of Zhang, H
	Transliteration of (Asian, Cyrillic, etc.) names	Hsin vs. Sin vs. Xin
Lack of journal title standardization	Various abbreviations and punctuation styles in journal titles	Heteroatom Chemistry vs. Heteroat. Chem. vs. Heteroatom Chem
Numeric bibliographic fields (publication year, volume number, pagination)	Transposed digits	p. 564 vs. p. 654
	Plus or minus one digit	1997 vs. 1998

Error rates, defined as citing references with discrepancies which resulted in a non-match in WoS, are reported to be 6.2% (Larsen et al., 2007), 7% (Tunger et al., 2010), 9.4% (Moed & Vriens, 1989) and 12% (Hildebrandt & Larsen, 2008). Moed (2005) carried out the most comprehensive study on the accuracy of citing references in WoS. He investigated 22 million citing references by employing different match keys in order to match the references to their 18 million target articles. He found 7.7% discrepant references, which resulted in a non-match in WoS, i.e. a missed citation. However, as discussed in section 3.4, the definitions of an error, discrepancy and inaccuracy differ in all these studies. The error rates are, therefore, not strictly

comparable, but still permit an estimate that the average missed citation rate (MCR) in WoS may range between 6 and 12%.

4.2.1 Author names

Author names and their variations, different spellings and inconsistencies constitute the most frequently discussed aspects in previous research. On the one hand, the problem areas deal with technical differences on a typographical level, such as names with diacritics or special characters, due to the fact that such formats are not supported in WoS and GS[16] because the values are decoded in ASCII, double middle initials with or without punctuation, and misspelled author names with adjacent consonants or ligatures on account of OCR errors (Harzing, 2008; Meho & Yang, 2007; Tunger et al., 2010). On the other hand, some of the variations have a stronger semantic influence, as they could also stand for two different authors. We can differentiate between two types of semantic variations: one is compounded names, which can be prefixed with a foreign article, hyphenated or consist of several parts. Typical examples of multiple-part last names are found in Spanish, Portuguese, Indian and Asian names (Garfield, 1981; Hood & Wilson, 2003; Meho & Yang, 2007). The other type is transliterated names from a non-Latin alphabet, such as the Cyrillic or Arabic alphabet (Garfield, 1990). Additionally, incorrect first initials caused by the use of nick names are another semantic challenge in the correct matching of author names.

4.2.2 Publication names

Bibliographic references contain different kinds of journal variations and abbreviations (Hood & Wilson, 2003; Jacsó, 2008d). Ideally, citing authors would use the full publication name or the ISO abbreviation for an easier match in the bibliometric database, but this is not always the case. Hence, it is the task of the bibliometric data source to detect and consolidate different variations of the same journal title. While Reedijk (1998) and Harzing (2008) criticize WoS for its poor aggregation of minor journal variations, Jacsó (2006) and Franceschet (2010a) report that journal title normalization in Scopus and GS works less well than in WoS. A short contribution on the SIGMETRICS mailing list[17] (August 8, 2013) entitled "Problems with Web of Science" also discussed the lack of journal standardization in WoS and the resulting missed citations. Two well-known researchers in the field of bibliometrics, namely Y. Gingras

[16] Scopus can deal with accented and special characters in its search and retrieves results containing both variants. It also matches Greek characters and finds common American/British English variant spellings. (Elsevier B.V., 2014b)

[17] http://listserv.utk.edu/cgi-bin/wa?A2=sigmetrics;h0YNGQ;20130808093255-0400

and L. Leydesdorff, commented on the question resignedly that the lack of journal title normalization could be obviated given an appropriate search strategy.

4.2.3 Numeric bibliographic fields

Inaccuracies in numeric bibliographic fields have been described as either missing or wrong volume numbers and starting pages. More specifically, the digits in one field or the values of entire numeric fields may have been swapped or may differ in only one or two digits (Larsen et al., 2007). Incorrect volume numbers may be related to dual-volume numbers or combined volumes (Moed, 2005). Differing page numbers can occur in the electronic and paper-copy versions of the same articles (Moed, 2005), but can also be related to a starting page number bearing the cited page number (Larsen et al., 2007). For example, in US law journals it is "normal" to cite the page number of the quote instead of the first page number (Moed, 2005). This phenomenon is, in general, more common in the SSH than in the NS. Analogously to the other inaccuracies described in this section, inaccuracies can derive from an author's inattention, from the editorial conventions of a journal or a required citation style, as well as from data capturing and formatting procedures at WoS.

4.3 Summary

This chapter discussed data accuracy and inaccuracies in bibliometric data sources. We illustrated how the accuracy of one bibliometric data source, i.e. the bibliographic references of a publication, can influence the accuracy of the other two, i.e. the citation indexes and the databases of the applied bibliometric research groups. While other aspects, such as the choice of bibliometric indicators, the inclusion of certain document types, the choice of citation windows, etc., impact the results of bibliometric studies in general, the very foundation of accuracy is the data values of bibliographic references. The accuracy of these data values is also influenced by the processes of data extraction, cleaning and harmonization as well as by the citation matching process, which may either correct existing inaccuracies or introduce additional ones.

In our data analysis, we concentrate on inaccuracies in references in the fields defining bibliographic data accuracy, as described in section 3.4. Additionally, they coincide with the bibliographic fields typically used in citation matching: first author name, publication year, publication name, volume number and starting page. The problem areas of these bibliographic

fields, such as errors, variations and inconsistencies in author names, publication names and numeric bibliographic fields, discussed in section 4.2, impacted the set-up of the coding scheme in chapter 6. Due to time and resource restrictions, semantic challenges, such as author name disambiguation and the differentiation of publications in the sense of FRBR, are not further analyzed in this research.

5 METHODOLOGY

Several issues relevant to the methodology of this research have been discussed in the previous sections on Citation matching (section 2.2.3), Bibliometric data sources (section 2.4), Data accuracy assessment (section 3.3), Bibliographic data accuracy assessment (section 3.4) and Inaccuracies in bibliographic data values with a primary impact on the citation matching process (section 4.2). This chapter presents the methodology eventually applied to answer the research questions posed in this doctoral research. We first give an overview of the terminology used in the evaluation (section 5.1), discuss the research methodology (sections 5.2 and 5.3), explain the data sampling process (sections 5.4 and 5.5) and conclude by describing the data collection process (section 5.6). Section 5.7 summarizes the chapter.

WoS was chosen as the baseline for the evaluation in this research because it offers the *Cited Reference Search* feature, which enables one to search for missed citations that could not be matched automatically by the WoS citation matching algorithm (cf. section 2.2.4). Neither Scopus nor GS provides this functionality. We used the web versions of all three databases, as they are the typical entry points available to most users (Ball & Tunger, 2006).

5.1 Definition of terminology for the evaluation

As bibliometricians differentiate between cited or target articles and citing or source articles (cf. section 2.2.1 on Cited and citing articles), this wording has been adopted and interpreted for the evaluation in this dissertation. Table 5 summarizes the terminology: the WoS records of cited articles are assessed against the bibliographic data from the original articles (PDF or paper version). Both are referred to as target articles (cf. first two rows in the target articles section). Next, the references from the citing articles (PDF or paper version) are assessed against the bibliographic data from the WoS records and the original cited articles[18]. The citing

[18] Due to cost-efficiency reasons, original cited and citing articles have not been used in citation matching processes in previous research. However, this thesis aims to cover all possible sources of inaccuracy and, therefore, the effort was made to manually collect all original cited and citing articles.

articles are referred to as source articles. Consequently, all bibliographic fields from the target articles are referred to as target data fields and those from the source articles as source data fields (cf. first two rows in the source articles section). Likewise, instances of data fields are called data values and can either be target data values (from the original article or WoS record) or source data values (from the reference of the citing article). A data record consists of different data fields that each holds one or more data values. One data record holds information about one cited article. Last, the cited reference information from all missed citations in WoS and Scopus (cf. second two rows in the source articles section) is assessed against the cited reference information from a correctly matched citation for the respective cited article (cf. second two rows in the target articles section). They are referred to as CitRefmatch target records, holding the correct cited reference information, and as CitRefmiss source records, holding cited reference information from a missed citation.

Table 5: Terminology of the data assessment process

	Origin of data	Data record	Data field	Data value
Target article	Original article (PDF, paper version)	Original target record	Original target data field	Original target data value
	WoS record	WoS target record	WoS target data field	WoS target data value
	Citing reference in WoS - matched	CitRefmatch-WoS target record	CitRefmatch-WoS target data field	CitRefmatch-WoS target data value
	Citing reference in Scopus - matched	CitRefmatch-Sco target record	CitRefmatch-Sco target data field	CitRefmatch-Sco target data value
Source article	Reference in citing article (PDF, paper version)	Source record	Source data field	Source data value
	Citing reference in WoS - missed	CitRefmiss-WoS source record	CitRefmiss-WoS source data field	CitRefmiss-WoS source data value
	Citing reference in Scopus - missed	CitRefmiss-Sco source record	CitRefmiss-Sco source data field	CitRefmiss-Sco source data value

According to the findings of the pre-study of bibliographic data accuracy (Olensky, 2012), the following bibliographic data fields are assessed in this doctoral research: *author names*, *first* and *second initials* of their first and second given names, *article title*, *publication name*, *volume number*, *publication year* as well as *starting and ending page*. The issue number was not part of the accuracy assessment. On the one hand, only a small number of studies investigated the issue number in the context of bibliographic data accuracy (cf. Table 3); on

the other hand, historically, the issue number was not made (and is still not) part of the coding in the cited reference information in WoS. For the bibliographic field *author name*, the assessment was divided into three data fields in order to obtain more accurate results: *author last name*, *first initial* and *second initial*. The subfields can contain more than one instance per record, since an article can have more than one author (cf. Figure 7). Therefore, the number of assessed author data values varies from record to record.

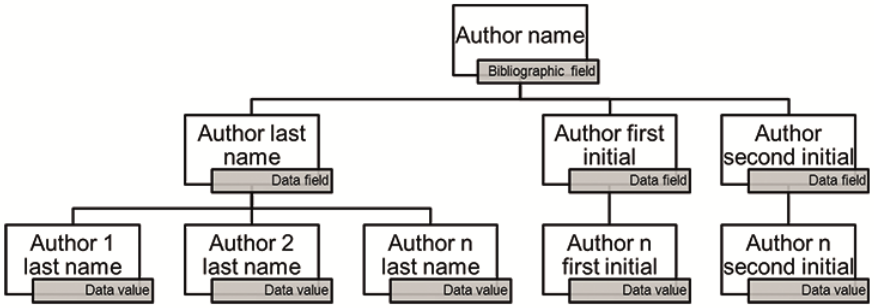


Figure 7: Levels and instances of the bibliographic field *author name*

5.2 Qualitative content analysis

This dissertation aims to convey a full understanding of the characteristics, patterns and causes of inaccurate bibliographic data that can influence the citation matching process. This was achieved by conducting an assessment of the data accuracy of citing references. On the one hand, such data quality assessment processes are carried out automatically in database management systems by different string matching methodologies (cf. section 2.2.3 and 3.3). The Levenshtein distance function, tested for the purpose of this research (Olensky, 2013; cf. section 3.4), does not reflect the severity of inaccuracies in bibliographic data, but it can be used to detect them. On the other hand, the accuracy of references in bibliometric databases has been investigated by applying match keys analogously to the citation matching process (cf. section 2.2.3). This method does allow a larger number of records to be investigated, but it does not pinpoint which inaccuracies the algorithm was able to handle. In our research question RQ1, we ask what types of inaccuracies occur in a bibliometric data source and how they can be categorized. In other words, we aimed to acquire an in-depth understanding of the characteristics of inaccurate bibliographic data, which neither of the automatic assessment methods is able to provide. Thus, we explored qualitative research methods that would support this aim. To the best of our knowledge, the only qualitative research method that allows a

categorization of "recorded information sources" (Beck & Manuel, 2008, p. 18) is content analysis, a method which, therefore, was used in this doctoral research.

Content analysis is a "research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use" (Krippendorff, 2004, p. 18). Hence, content analysis is a method that not only works for text analysis, but can also be applied to any "identifiable message or message component", i.e. they "can be words, characters, themes […]" (Neuendorf, 2002, p. 71). It has been applied to various contexts, both scholarly and commercial (Neuendorf, 2002). The most commonly known examples of applications are the analysis of interviews containing open-ended responses, of media content (e.g. violence on TV or specific topics on the news) or of larger numbers of texts in linguistics to study the syntax, semantics or style (Neuendorf, 2002). As the subject of our investigation is bibliographic data records, i.e. "*objets trouvés*, ready-made material existing to hand", these "records […] can be subjected to content analysis" (Slater, 1990, p. 122). Typically, content analysis of pre-existing, structured material (as opposed to data material gathered in interviews or observational studies and then subjected to content analysis) is a qualitative research method in historical research (Slater, 1990), but it has been applied in the field of library and information science as well: Cronin (1982), Lynch & Smith (2001) and Croneis & Henderson (2002) have all analyzed job advertisements; Haas & Grams (2000) have analyzed web pages and links contained therein; Turner & Beck (2002) have applied content analysis to code repair strategies of users searching online catalogues; and Marsh & White (2003) have developed a thesaurus of image-text relationships.

The method seeks to derive generalizable conclusions from the units of analysis. It employs a coding form to extract information about pre-defined variables from the message units and a codebook to categorize the extracted messages (Neuendorf, 2002). Content analysis can be used qualitatively, if the establishment of the coding scheme is part of the process, or quantitatively, if the analysis requires an a priori design, i.e. the codebook and the coding form must be constructed in advance (Neuendorf, 2002; Schreier, 2012). To the best of our knowledge, content analysis has not been applied to bibliographic data before, therefore, we developed a methodological framework that supports the specific requirements of analyzing bibliographic data and employs qualitative content analysis (Figure 8). We conduct an automatic assessment of the data (Part A) which prepares the units of analysis, i.e. data values from the different bibliographic fields, for the qualitative content analysis (Part B). The complementary quantitative analysis (Part C) evaluates the frequency of inaccuracies. Part D describes the evaluation of missed citations. In this evaluation, the abilities of matching

algorithms of five other data sources were compared by means of the WoS missed citations. Moreover, we investigated which inaccuracies caused the non-match in WoS and triangulated the data with the results of the other bibliometric data sources.

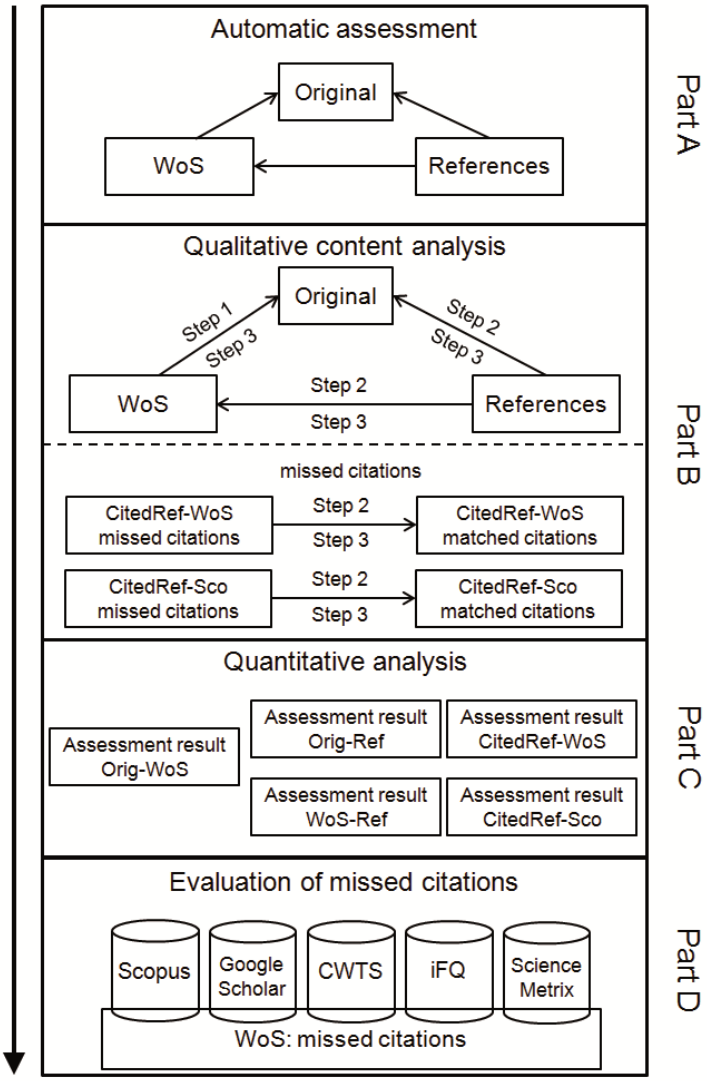


Figure 8: Qualitative content analysis adapted to bibliographic data assessment

First, we automatically assessed the data with the Levenshtein distance function (Figure 8, Part A) to detect data values containing discrepancies (Olensky, 2013; cf. section 3.4). Part B continued by intellectually assessing the discrepant values. The intellectual assessment represents the qualitative content analysis and followed the steps defined by Neuendorf (2002) by establishing a codebook, a coding form and the coding of the inaccuracies (cf. chapter 6). The goal was to record in detail the requirements of converting a discrepant value into the correct one that could eventually be implemented as rules in citation matching algorithms. The

intellectual assessment consisted of three steps. In the first step, the bibliographic data of the original target article was assessed against the bibliographic records from WoS by scrutinizing each discrepancy found, identifying, if possible, the cause of the inaccuracy and setting up a basic coding scheme of inaccuracies (e.g. B = spelling error, O = incorrect order of authors). In the second step, the bibliographic data from the source articles was assessed against the original target records as well as the WoS target records to assign the codes to the inaccuracies and further edit and expand the coding scheme. Moreover, the cited reference information of the missed citations identified was assessed against the cited reference information of correctly matched citations (for WoS and Scopus). The third step entailed two intellectual assessment iterations in which all four assessment processes were repeated. The first iteration checked the data values and the assigned inaccuracy codes (IACs) for consistency. The coding scheme was further edited and streamlined. The second iteration resulted in the final assignment of the inaccuracy codes. The assessment results are coded as depicted in Part C of Figure 8:

- WoS records against original articles: assessment result Orig-WoS
- References (source records) against original articles: assessment result Orig-Ref
- References (source records) against WoS records: assessment result WoS-Ref
- Cited reference information of missed and matched citations in WoS: assessment result CitedRef-WoS
- Cited reference information of missed and matched citations in Scopus: assessment result CitedRef-Sco

We carried out the content analysis ourselves and, therefore, took some validity and reliability measures (Gibbs, 2007, p. 96f.):

- Data was checked several times for completeness and consistency.
- It was ensured that there was no shift in the meaning of the codes during the process of coding by allowing several weeks to elapse between the assessment cycles and the consistency checks. The coding process strictly followed the coding procedure and codebook explained in chapter 6.
- The codebook was peer reviewed by colleagues.
- Inaccuracies were constantly compared with each other to check the consistency and accuracy of the codes and their application.
- Codes were cross-checked twice (two final assessment iterations in Part B-Step 3; cf. Figure 8).

On completion of the qualitative assessment, a quantitative analysis of inaccuracy codes was conducted, providing a statistical distribution of inaccuracies in the references of the citing articles over the different facets of the data sample (Part C). The facets used for evaluation are: domain of the cited articles, discipline of the cited articles, language of the cited articles, language of the citing articles, publication year of the citing articles (summarized into three five-year citation windows) and document type of the citing articles. The results of Part C are discussed in chapter 7 for the three assessment results Orig-WoS, Orig-Ref and WoS-Ref. The assessment results CitedRef-WoS and CitedRef-Sco are discussed in Chapter 8.

Finally, Part D entailed the evaluation and comparison of citations missed in WoS as processed in the five other bibliometric data sources. On the one hand, we compared how many of the citations missed in WoS were covered and matched by the citation indexes, Scopus and GS, as well as by three applied bibliometric research groups, CWTS, iFQ and Science-Metrix. On the other hand, to answer RQ3, we triangulated the results of what types of inaccuracies caused missed citations in all six bibliometric data sources. Triangulation renders results more accurate and credible by applying different approaches to the research problem (Patton, 1999). One can triangulate data (from different sources), investigators, theories or methodologies to "*situationally* check the accuracy and repeatability of the specimens and emerging causal proposition" (Denzin, 1989, p. 93). As the subject of our investigation is data accuracy in bibliometric data sources, it was logical to triangulate different data sources. Data triangulation allows us to determine what kinds of inaccuracies impact the citation matching process with greater confidence. As explained in section 2.4.3, only the data from the research group CWTS is used exactly as-is in citation analyses for research assessment. Science-Metrix uses Scopus raw citation data complementarily and iFQ's algorithm is not in production yet. Hence, missed citations in the CWTS database are more significant than the missed citations in the other data sources.

5.3 Assessment of variants

Two of the bibliographic data fields, *publication name* and *article title*, were evaluated using two different variants in the assessment processes of source records against original articles (Orig-Ref) and WoS records (WoS-Ref). In this section we explain the specifics of these processes. Figure 9 illustrates the assessment process for the bibliographic field *publication name*. The publication names from the source records were assessed against the full publication name (the corresponding data field in WoS is *SO, Publication Name*) as well as the

ISO abbreviation of journal titles as recorded by WoS (the corresponding data field in WoS is *JI, ISO Source Abbreviation*), which resulted in two variants of the assessment result WoS-Ref. These variants were consolidated into one assessment result prior to the quantitative analysis. Section 7.9 discusses the consolidation and evaluation of the variants in more detail. The original target articles did not contain any variants of the *publication name*.

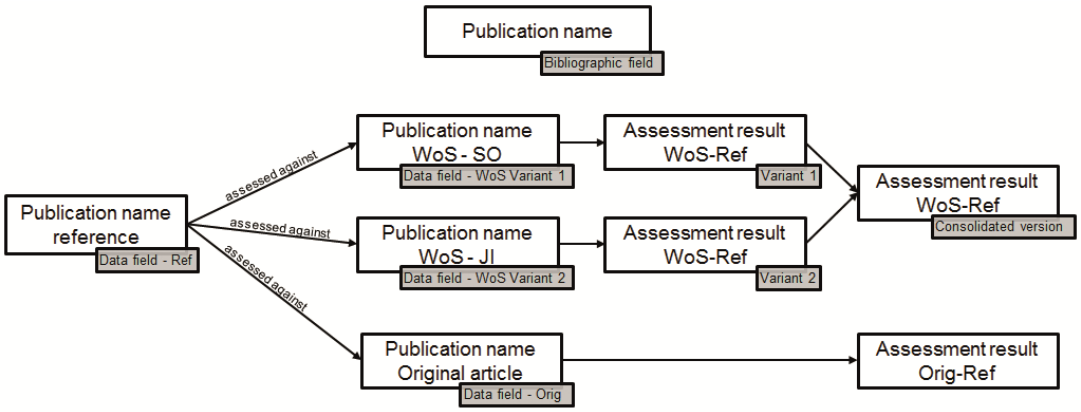


Figure 9: Assessment process for the variant *publication name*

The bibliographic field *article title* had an optional variant stemming from the references: *Translated article title* (cf. Figure 10). Some of the source records (from the German dataset) gave an additional translation of the article title in brackets, which was spotted during the data entry process and turned into an additional opportunity for analysis. Interestingly, the English translations could either be found in brackets after the original German article title or they were cited as if they were the original article title and the original German article title was given in brackets. The two parts were separated from each other in the data parsing process in order to assess which part of the title was the translation and which provided the most accurate results. The two variants of the *article title* were then assessed against the target values and resulted in two assessment results for the bibliographic field *article title* in both assessment results. Analogously to the above described consolidation of the *publication name*, the two variants were consolidated into one prior to the quantitative analysis.

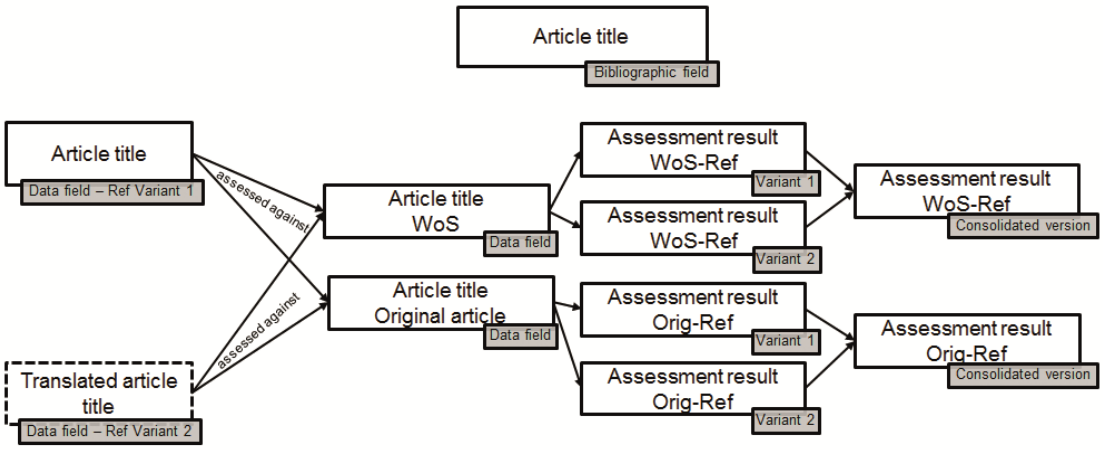


Figure 10: Assessment process for the variant *article title*

5.4 Stratified purposeful sampling

Since the total number of records in WoS exceeds 46 million, it was not possible to carry out a statistically representative quantitative or qualitative study of the accuracy of references with the resources given in this doctoral research. Most inaccuracy studies (e.g. Moed, 2005; Franceschini, Maisano & Mastrogiacomo, 2013a, 2013b) state that the distribution of errors in references is highly skewed and does not follow any specific patterns. Therefore, this research aims to cover a data sample that represents a sub-universe of typical characteristics of publications and bibliometric data sources, such as different languages, document types or scientific disciplines, which can influence the calculation of bibliometric indicators (Moed, 1996). In the context of bibliometric analyses, data sources are often discussed and compared with others in terms of their different facets of coverage: format, temporal and geospatial coverage as well as domain and discipline (e.g. Bakkalbasi et al., 2006; Falagas et al., 2008; Meho & Yang, 2007; cf. section 2.4.2). Thus, we interpreted these five facets (domain, discipline, document type, language and publication year) as the baseline for selecting the cited articles in our data sample. From the experience of previous studies (Larsen et al., 2007; Hildebrandt & Larsen, 2008), we aimed for a sample that would cite typical cases of publications in WoS and would be of a size around 3,500 to 4,000 citations that could be handled with the given resources.

Typical cases of a population can be selected by applying a stratified purposeful sampling approach (Patton, 2002). Purposeful sampling allows one to select information-rich cases, stratifying the sample means selecting different samples from a larger population according to

different characteristics; samples can be nested (Patton, 2002). We combined two types of purposeful sampling as described by Patton (2002): on the one hand, we selected cases typical of the total population, i.e. publications in WoS, as our starting points (the cited articles); on the other hand, we opted for homogeneous data samples describing particular sub-groups in-depth, i.e. citations which cite these typical cases (e.g. all citations within a specific citation window or all citations citing an English article). The citing articles were not subject to any restrictions and were summarized into homogeneous groups for the quantitative analysis.

In the stratified purposeful sampling process the following decisions were taken to determine the different strata of the cited articles, i.e. target articles.

Domain. WoS indexes journals in the NS as well as in the SSH. Both domains should be represented as typical cases of publications in WoS.

Discipline. Following the study of Larsen et al. (2007), we determined that three different disciplines within the two domains, NS and SSH, should be represented. How these six disciplines were selected will be explained in section 5.5.

Language. The next selection criterion was the language of the cited articles: English, since the majority of articles in WoS are in English and they represent the majority of typical cases in WoS, and German, the author's mother tongue and a language that contains typical sources of inaccuracy (e.g. German Umlaut; cf. section 4.2), were chosen.

Document type. Due to the different classifications of documents in data sources (as described in section 2.4.2), we elected only to work with articles, as classified by WoS, even though this might mean missing a few misclassified documents with potentially higher citation counts. Yet, the selection was purposely made to further narrow down the data sample and to investigate a typical document type in WoS, which would also typically be used in citation analysis. The inclusion of the citing articles in the data sample was not restricted to any specific document type.

Publication year. 2003 was selected as the first publication year, giving the articles a 10-year citation window from the current year (2012) to accumulate a reasonable number of citations. 1998, i.e. 5 years earlier, was chosen as the second publication year in order to study whether the inaccuracy patterns change over time. Furthermore, it allows one to investigate whether the citation matching algorithms have changed over time, i.e. whether they have kept pace with technological advances.

The domain, discipline and language of the cited articles were used as facets in the quantitative evaluation (cf. sections 7.3, 7.4 and 7.5), whereas the other two facets, document type and publication year of the cited article, were merely used to extract typical cases from the total population and limit the size of the data sample.

5.5 Data sample

The actual data selection process (Figure 11) commenced by identifying German-language journals indexed in WoS with the help of the Journal Citation Report 2011[19]. In general, German-language articles do not have high citation rates in WoS; that is why we decided to work with the top 10 cited articles from each German-language journal from each publication year (1998 and 2003). All journals from the JCR Science Edition 2011 that were classified under the countries Germany, Austria and Switzerland were selected to determine the three journals in the NS. A total of 257 Austrian, German and Swiss Journals is listed in the JCR Science 2011, of which 165 gave Multi as their language, 89 German, 2 English and 1 French. Journals with mixed languages were excluded and, of the remaining 87 journals[20], all top 10 cited articles were searched for the years 1998, 2003.

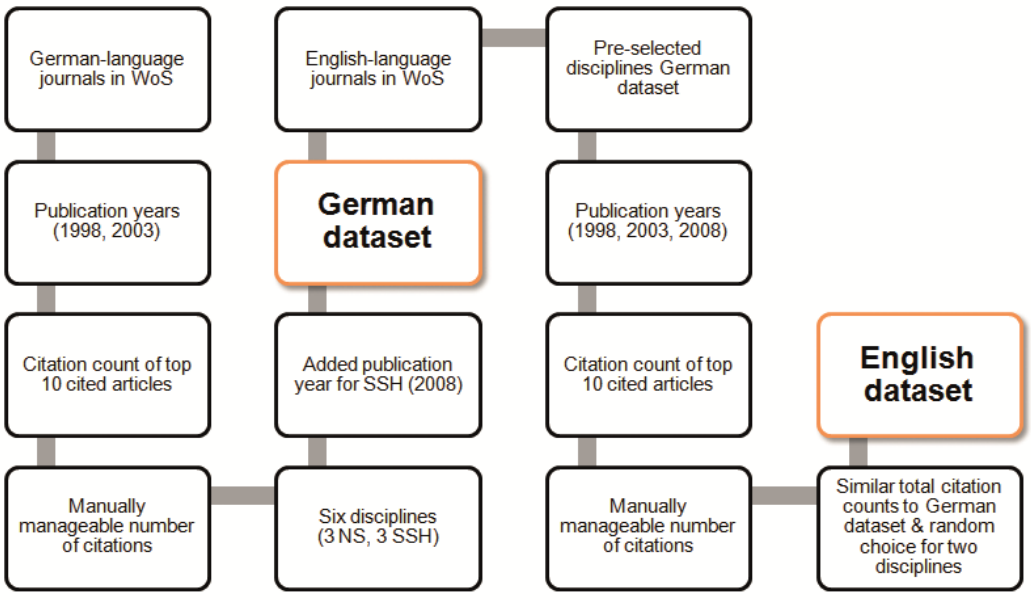


Figure 11: Selection process of the data sample

[19] The Journal Citation Report 2011 (JCR) had to be used because, at the time of starting the data sample selection, the JCR 2012 was not yet available.

[20] Two German-language journals actually only contained English articles and were therefore excluded.

All journals that did not provide coverage of at least 10 articles in 1998 and 2003 were excluded. Additionally, journals that were not covered in Scopus were excluded as well. Of the remaining 15 journals, three journals were selected that provided a balance between a "sufficiently high" and "still manually manageable" number of citations. The threshold was defined as below 600, but above 300 for each journal. For the German data sample this meant simply selecting the three journals with the highest citation counts for the two publication years. Since, of the last 15 journals, 10 were classified as medical journals, we chose two medical journals, but opted for two different medical fields: *Der Orthopäde* (WoS subject category: Orthopedics); *Deutsche Medizinische Wochenschrift* (WoS subject category: General & Internal Medicine), and one chemical journal: *Chemie in unserer Zeit* (WoS subject category: Multidisciplinary Chemistry and Chemical Engineering) [21].

To select three journals from the SSH, the same selection procedure was employed. All journals from the JCR Social Science Edition 2011 that were listed under the countries Germany, Austria and Switzerland were retrieved. The total number of Austrian, German and Swiss Journals in the JCR Social Science 2011 is 152, of which 25 gave Multi as their language, 75 English, 50 German and 2 French. We excluded any journals that did not provide the required coverage of articles in 1998 and 2003 (at least 10 per year). Then the journals were ranked, based on the total citation count of the top 10 cited articles in 1998 and 2003. The number of citations of articles in German-language SSH journals was even lower than in NS journals. The highest citation rates of typical SSH disciplines [22] were found for the following three journals, thus eliminating the need for further selection: *Berliner Journal für Soziologie* (WoS subject category: Sociology); *Politische Vierteljahresschrift* (WoS subject category: Political Science); *Zeitschrift für Pädagogik* (WoS subject category: Education & Educational Research). In total, the top 10 cited articles from these three journals were only cited by 371 references in 1998 and 2003. For this reason, an additional publication year (2008) was included to stock up the data sample, increasing the total number of source articles to 520.

Since the disciplines in the NS and SSH had already been predefined by the German-language journals, all English-language journals in those six disciplines were selected (as defined in the

[21] We are aware that the WoS subject categories are a controversial subject of discussion and not suitable for comparison with other data sources. However, for the purpose of selecting the initial 300 target articles, we regarded the subject categorization as sufficient.

[22] Psychology and Psychiatry were excluded from the selection. Both WoS and Scopus consider them as part of the social sciences, but usually they tend to be associated with Medicine and are, therefore, not considered as typical social sciences.

JCR 2011[23]). English-language journals were defined as those that gave English as their language and were assigned to one of the following countries: Australia, Canada, England, Ireland, New Zealand, Scotland, USA and Wales. Journals that either did not match the coverage criteria of at least 10 articles or were not covered in Scopus were excluded. For the remaining journals, the citation counts of the top 10 cited articles for 1998 and 2003 were retrieved. This process was repeated for all six disciplines, and for the SSH disciplines the citation numbers for 2008 were also recorded.

18 out of 154 journals in the subject category Multidisciplinary Chemistry matched these criteria, as did 7 out of 65 Orthopedics journals, 11 out of 155 General & Internal Medicine journals, 35 out of 206 Education & Educational Research journals, 19 out of 149 Political Science journals and 15 out of 138 Sociology journals. Since more than one journal per discipline met the criterion of having a manually manageable number of citations (below 600, more than 300), we chose the journals with the most similar total citation counts compared to their German counterparts to compile a data sample as balanced and homogeneous as possible. For two disciplines (Sociology and Education & Educational Research) in which more journals met this criterion, a random choice was made. Finally, the following six journals were added to the data sample: *Hand Clinics* (WoS subject category: Orthopedics); *Journal of Travel Medicine* (WoS subject category: General & Internal Medicine); *Heteroatom Chemistry* (WoS subject category: Multidisciplinary Chemistry), *Sociological Inquiry* (WoS subject category: Sociology); *Political Theory* (WoS subject category: Political Science); *Journal of Curriculum Studies* (WoS subject category: Education & Educational Research).

The resulting data sample consists of 300 cited articles[24] from 12 different journals covering six different disciplines from three different publication years plus their corresponding citing articles within three variable five-year citation windows (1998-2002, 2003-2007 and 2008-2012). The citing articles are not subject to any further restrictions. Thus, in the quantitative analysis of bibliographic inaccuracies (chapter 7), we evaluated the inaccuracies based on the domains, disciplines and languages of the cited articles as well as the document types, languages and publication years (i.e. citation windows) of the citing articles. The strata covered in the data sample for the cited articles overlap; therefore, sampling units can belong to more than one stratum, as the illustration in Figure 12 shows. The cited articles cover the NS and the SSH; within each domain three different disciplines are represented. For each

[23] We are aware that subject categories may vary in different years and different JCRs, respectively. We relied on the subject category information from the year 2011.
[24] The list of all cited articles is given in Appendix B.

discipline we selected two journals, one in English and one in German. Each journal provides the top 10 cited articles from two (three for the SSH) publication years. Since only the inaccuracies in the citations to the cited articles are compared, the additional publication year for SSH cited articles allows those articles to accumulate more references and, therefore, reflects a more complete picture of occurring inaccuracies.

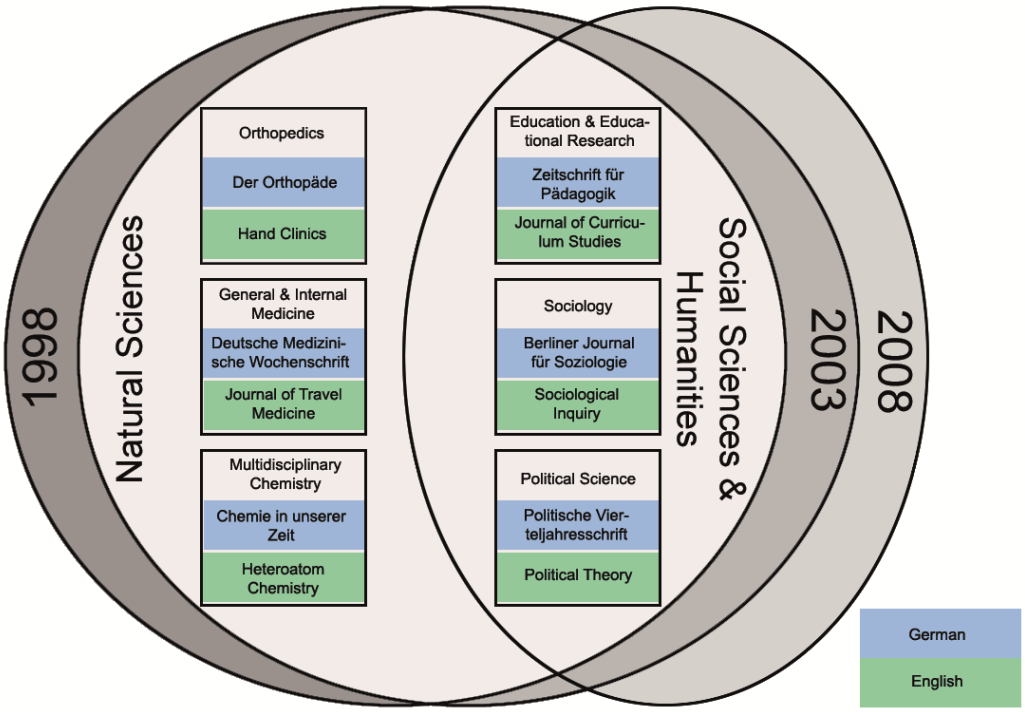


Figure 12: Strata of the data sample (cited articles)

5.6 Data collection

The data collection started with conducting a *Cited Reference Search* for all 12 datasets. One dataset consists of all cited articles from one journal (20 in the NS and 30 in the SSH) and their corresponding citing articles. As the *Cited Reference Search* allows searching for permutations of author names, journal titles and publication years to identify missed citations, we combined these bibliographic fields with each other in different ways and made use of wildcards. For example, the cited article *Leitbild ist nicht mehr der erwerbstätige, sondern der tätige Mensch* by *Bernd Zymek* in the journal *Zeitschrift für Pädagogik* (publication year: 1998; volume number: 44) was searched by combining different spellings of the author's last name

with different spellings of the journal as well as different publication years (cf. Table 6)[25]. As soon as we found a potential missed citation, we downloaded the information of the citing article and added it as possible citation to the dataset. The verification, whether the citing article truly contained a citation to the cited article, was carried out in the data entry.

Table 6: Example of combinations in the *Cited Reference Search*

Journal	Author name	Publication year
z* pa*	Z*mek	1998
z* pad*	C*mek	199*
z* paed*	Zim*k	189*
z* ped*	Zym*k	198*
***journal variation* plus**	C*mek	
in press	Cim*k	
inpress	Cy*k	
in druck	Zy*k	
indruck	Zy*	

The full bibliographic records of all articles (cited and citing articles, including those from the *Cited Reference Search*) for all 12 journals were downloaded from WoS between February 2013 and August 2013. A total number of 3,992 citing references was found for the 300 citing articles in WoS. The records were stored as MS Excel files as well as in a MySQL database. Each cited and citing article was given a unique ID, which was an important identification tool for the data entry process and the data analysis. The articles were retrieved from the university libraries of Humboldt-Universität zu Berlin and National Taiwan University, either online or as a scan of the paper version of the journal. Publications that were not available were ordered via interlibrary loan at Humboldt-Universität zu Berlin. The data collection of the articles started in February 2013 and ended in January 2014. We identified 4 duplicate records and 33 false positives (2 of the latter being corrections, that are listed as citations in WoS), which we excluded from the analysis. False positives are citing articles which were matched to an incorrect cited article by WoS and, therefore, inflate its citation count. 26 citing articles (one of them a citing article with a missed citation) were not obtainable, therefore, we also had to exclude them from the analysis. In total, the data sample covers 3,929 verifiable references from 3,735 citing articles.

The comparative search for missed citations and the download of the respective data in Scopus and GS were completed between December 8 and 19, 2013. Each cited article containing citations missed by WoS was searched in Scopus and GS and the bibliographic data of cited and citing articles were downloaded. The records of the citing articles that were also not

[25] Please refer to Appendix C for a detailed description of the search strategy in the *Cited Reference Search*.

matched by Scopus and GS (compared to the total number of 219 missed WoS citations), but were covered in the database, were also downloaded. As already reported by other studies (e.g. Meho & Yang, 2007), the search and download in GS took significantly longer than in Scopus, since GS does not provide an interface for downloading records. Additionally, GS complicated the download by assuming we were an automatic program that was illicitly downloading records. Hence, we had to switch country sites (http://scholar.google.de, .fr, .es, .com) and reset the browser settings several times. Even though Publish or Perish is a software tool that helps with searching, organizing and de-duplicating citation counts in GS (Harzing, 2008), one cannot access and download the citing articles with this tool to further analyze the citations. However, in this research, access to the citing articles was crucial to determine whether references had been covered and correctly matched by GS. For this reason, the GS web access was used.

We received the data from the three applied bibliometric research groups between February and April 2014. We provided them the bibliographic data of the 300 cited articles and asked them to run these through their databases, applying their citation matching algorithms. Additionally, they checked whether the missed WoS citations were indexed in their databases in order to exclude non-matches, due to lack of coverage, from the analysis.

Data entry. Data entry of the bibliographic data from the original articles (cited and citing) as well as the manual extraction of the cited reference information from the bibliographic source records (CitRefmatch/miss from WoS and Scopus) was handled by two student assistants who worked independently of each other on the same datasets. The datasets were compiled according to the journals and consisted of a Microsoft Excel file for data entry, the articles as PDF files and the bibliographic records of the missed citing articles (txt files). The first task of the student assistants was to collect, and enter into the first sheet, the bibliographic data of the cited article: *author's last name*, *first initial*, *second initial*, *article title*, *publication name*, *volume number*, *starting* and *ending page* as well as the language of the article. As mentioned in section 5.1, the issue number was not collected. Second, they went through the citing articles, looked for the reference of the corresponding cited article and entered the data in the Excel file exactly as given in the articles' bibliographies including existing discrepancies, punctuation and information that might not be directly attributable to one of the bibliographic data fields. During this task, potential missed citations were verified. Last, they recorded the cited reference information to the cited, but not-matched, articles as well as an example of cited reference information from a correctly matched reference for both, WoS and Scopus. (cf. Figure 4 and Figure 5 for an example from WoS). For the evaluation of the inaccuracies in

citations missed by GS, the assessment results of the original reference (Orig-Ref) were used, since we assumed that GS scans the indexed documents and extracts citation data directly from the original document. For the evaluation of the inaccuracies in citations missed by the applied bibliometric research groups, the assessment results of the CitedRef-WoS sample were used.

Data handling. The bibliographic data gathered by the student assistants was first matched against each other to guarantee that the correct data had been recorded. If the data contained any discrepancies, recourse was made to the original records to examine their accuracy. During this process, the records were also checked for, and marked as, false positives (cf. section 7.10) and duplicate records. Afterwards, the data records were checked for completeness.

For the automatic accuracy assessment of the target and source articles, all punctuation (cf. Appendix D, Table 38) was replaced by space characters and multiple neighboring space characters were eliminated. We also cropped any full first or second given names and only recorded the initials. For the cited reference information, we left the data exactly as-is, since we could not exclude the possibility that, for instance, differing punctuation might determine a match or non-match. As the Levenshtein distance function is not a standard MySQL functionality, it was programmed and added manually (cf. Appendix D). Only inaccuracies that the Levenshtein distance function can detect on a technical level were considered. This excludes capitalization as well as 22 out of 91 tested special characters that are part of the subset Latin1-Supplement in Microsoft Office Word: e.g. é, è, ñ (all tested characters are listed in Appendix D, Table 39, the special characters which the LDF cannot detect are listed in Table 40). The intellectual assessment, i.e. qualitative content analysis, is described in chapter 6.

5.7 Summary

In this chapter we have introduced the methodology applied in this doctoral research. We adapted the research method *content analysis* to the specific requirements of analyzing bibliographic data by combining an automated assessment with a qualitative content analysis. We chose a stratified purposeful sampling approach to select a data sample according to characteristics typical of publications in WoS, which allows us to compare the results of homogeneous subgroups in the quantitative analysis (e.g. all inaccuracies in references to

English articles). Furthermore, we described the data collection process in WoS (including the *Cited Reference Search*), Scopus, GS and the applied bibliometric research groups as well as the data entry and handling procedures.

The methodology applied in this research answers different parts of the research questions: the combination of the automatic and the intellectual assessment (Part A and B), applied to identify and categorize bibliographic inaccuracies, answers RQ1. The results of the qualitative content analysis answer the first subquestion of RQ1, by organizing them into a taxonomy (cf. chapter 6). The quantitative analysis of the inaccuracy categories (Part C) answers the second and third subquestions of RQ1 concerning the frequency of occurrences and whether inaccuracies can be specifically related to one of the strata in the data sample. This part of the evaluation is discussed in chapter 7. The evaluation of the sampling units that were identified through the *Cited Reference Search*, i.e. missed citations, and the triangulation of multiple data sources (Part D) answer RQ2 as to whether specific categories of inaccuracies automatically lead to a missed citation. Chapter 8 discusses the results of the missed citation evaluation. RQ3 is answered by taking into account the answers to the previous research questions and inferring the type of data manipulation rules that need to be considered to further improve citation matching processes. Chapter 9 elaborates potential improvements to citation matching.

6 Constructing a Coding Scheme for Bibliographic Inaccuracies

This chapter describes the construction of the coding scheme for bibliographic inaccuracies which was set up in Part B of the analysis process and refers to the qualitative content analysis (Figure 8, p. 46). First, this chapter discusses the coding procedure (section 6.1) which describes a set of guidelines followed during the establishment of the codebook and the actual coding of the inaccuracies. The rules of the coding procedure ensured the objectivity and reliability of the results. Second, the codebook, listing all inaccuracy codes, is described (section 6.2). Section 6.3 organizes these inaccuracy codes into a taxonomy of inaccuracies in bibliographic references. The chapter concludes with a summary in section 6.4.

6.1 Coding procedure

This section describes the guidelines we followed in order to analyze the data objectively. The guidelines are based on the results of our study on the accuracy of references (Olensky, 2012) and of our pilot study on the assessment of bibliographic data accuracy (Olensky, 2013) as well as the theoretical chapters of this research. Logically, the assessment follows the definition of data inaccuracy from Chapter 3.2:

> *Data inaccuracy describes a discrepancy between the correct value v' and the assessed value v, i.e. any non-conformity between these two values is recorded. The term data inaccuracy is used synonymously with discrepancy. A data inaccuracy is not necessarily an error.*

The assessment direction. The assessment direction is graphically represented by the arrows in Figure 8 on p. 46. On the one hand, the WoS records were assessed against the original article records. In this case, the correct value *v'* refers to the target values from the original cited article and the assessed value *v* refers to the data values from the WoS records. On the

other hand, the source data values were assessed against the two datasets of target data values, i.e. the original (= cited) article and the WoS record. In that case, the correct value *v'* refers to the target data values from the cited article or the WoS record, whereas the value *v* refers to the source data value from the references in the citing articles. In the assessment process of the cited reference information from WoS and Scopus (CitedRef-WoS, CitedRef-Sco), the correct value *v'* refers to values from correctly matched citations and the value *v* to values from missed citations. Furthermore, even when the assessor knew that the target data value was incorrect (this was true for a few WoS records; cf. Appendix H), the assessment direction did not change. Hence, inaccuracy codes were also assigned to correct references.

Categorization of inaccuracies – coding form and codebook. During the assessment iterations, the codebook was established. The coding form was an Excel file used for the coding, providing a field for each assessed value. This allowed a description of each discrepant value individually, based on the error categorizations found in the study on the accuracy of references (Olensky, 2012). First, we identified whether the value was discrepant or missing and then described the nature of the discrepancy, such as two letters switched, different spelling, a missing digit, a number with only one incorrect digit, etc. After the first assessment cycle (Step 1 in Part B, cf. Figure 8) these broad categories were grouped together by identifying recurring patterns across bibliographic fields and clustering the inaccuracies. Based on these clusters, the inaccuracy codes were developed. We used a nominal scale consisting of letters, which had no meaning other than to label the categories. The coding scheme was set up with the technical implementation of data-handling rules in mind. When an inaccuracy was detected by the Levenshtein distance function, the goal of the qualitative content analysis was to find a way to convert the inaccurate value *v* into the correct value *v'*. The inaccuracy codes reflect discrepancies, ranging from fairly minor to more severe ones, for which no technical data manipulation rule could be identified.

Assignment of inaccuracy codes (IACs). The coding scheme consists of inaccuracy codes (IACs) labelled with capital letters and, for one IAC, with a combination of capital letters and numbers to track a specific aspect of the inaccuracy (cf. IAC *G*). If a data value contained different categories of inaccuracies, more than one category was assigned. Yet, if a value contained more inaccuracies of the same category, the inaccuracy did not multiply (Olensky, 2012). For instance, an IAC could be simply an *A* which translates into a spelling variation or could also be a combination of *A K* which translates into a *Typographical variation A* and a *Space* character discrepancy *K*. IACs in the data analysis were capitalized, listed in alphabetical order and had a space character in-between them. IACs were recorded in the

respective bibliographic data fields in five separate excel files for each assessment set (Orig-WoS, Orig-Ref, WoS-Ref, CitedRef-WoS, CitedRef-Sco).

Field independency. The goal of the assessment was to identify inaccuracy patterns that can be translated into machine-readable rules for data matching. Therefore, the identification, categorization and assignment of IACs relied as little as possible on the coder's background knowledge, but primarily endeavored to come up with IACs that could be translated into technically feasible rules. For instance, in Table 9 the publication name in the source data value is *Deutscher Medizinwochenschrift*. Even though we knew that the correct publication name was *Deutsche Medizinische Wochenschrift*, we assessed the values one by one and only compared the two data values in question. Consequently, the inaccuracies in the publication name of the source data value could only be detected in assessment no. 2 where *v'* is *Deutsche Medizinische Wochenschrift*, but not in assessment no. 1 (*v'* = Dtsch med Wschr) nor in assessment no. 3 (*v'* = Dtsch Med Wochenschr). However, we had to make a few exceptions. For example, to identify the correct citation of compounded names, we checked the author's own references to see how they cited themselves, assuming they would cite their own names correctly. Additionally, to detect a jumbled order of author names, it was necessary to consider other data values in the assessment of one particular data value.

6.2 The codebook

The final codebook consists of 32 different IACs with 25 main IACs and seven additional sub-IACs (*G1*-*G7*). The sub-IACs track which fields were interchanged with each other. Table 7 gives an overview of all IACs and indicates the main category of inaccuracies they belong to: *Type 1* summarizes all IACs that decode data fields that contain a correct value, but do not exactly match the entire correct value; *Type 2* refers to IACs that decode data fields that contain part(s) of a correct value; *Type 3* lists all IACs that decode data fields that do not contain a correct value. These three categories were used for the development of the taxonomy described in the following section 6.3. The IACs are listed and described in alphabetical order. For easier reference, the codebook can also be found in Appendix E (p. 204).

Table 7: The codebook

Inaccuracy code	Name	Type 1: contains a correct value	Type 2: contains part of a correct value	Type 3: does not contain a correct value
A	Typographical variation		x	
B	Spelling error		x	
C	Different language			x
D	Completely incorrect			x
E	Omitted			x
F	Cropped		x	
G	Interchanged fields	x		
G1	holds issue no	x		
G2	holds starting page	x		
G3	holds ending page	x		
G4	holds volume no	x		
G5	holds last name	x		
G6	holds first initial	x		
G7	holds second initial	x		
H	Jumbled value	x		
I	Abbreviation		x	
J	Partially incorrect		x	
K	Space	x		
L	Informational letter	x		
M	Incorrect interpretation of author names		x	
N	Additional information	x		
O	Incorrect order of authors	x		
P	No author name			x
Q	Special character		x	
R	Punctuation	x		
S	Padded	x		
T	Plus/Minus		x	
U	Full first name	x		
V	Incorrect interpretation of additional information		x	
X	Stop word		x	
Y	Word stem		x	
Z	Not available			x

Table 8: Overview of assessed data fields

Data field	Abbreviation used in the example tables	Type
Author last name Author first initial Author second initial	AuthorName[26]	String
Article title	ArticleTit	String
Translation of Article title	ArticleTitTrans	String
Publication name	PubName	String
ISO abbrev. Publication name	AbbrPubName	String
Volume number	Vol	String or Numerical
Publication year	PubY	Numerical
Starting page	SPage	Numerical
Ending page	EPage	Numerical

Before we discuss all IACs in the codebook, it is first necessary to explain the tables containing assessment examples in this section. Table 8 lists the assessed bibliographic fields and the corresponding abbreviations used in the example tables. With the help of Table 9, which gives an assessment example, this paragraph describes how to read them. The first row indicates the source data field (first column) and value (second column, in bold) from the citing article. This is the value that was assessed against the target data fields. For the example in Table 9, the data field is *PubName_Ref* (PubName = Publication Name; _Ref = from the reference of a citing article) and the value is *Deutscher Medizinwochenschrift*. The second row *Assessment no.* denotes an ID that was given to all assessments just for the sake of easier reference from the text to the assessment tables. The third row *Data field* indicates which bibliographic field is assessed, e.g. the author's last name, first or second initial, the article title, the publication name, etc. (as listed in Table 8) as well as the data source it comes from (_Orig = original article, _WoS = WoS record). The fourth row gives the actual target data value. Taking assessment no. 1 in Table 9 as an example, the data value *v'* for the data field *PubName_Orig* is *Dtsch med Wschr*. This means that one of the original articles held, as the publication name, the journal abbreviation *Dtsch med Wschr*, which is the target data value against which the source data values was assessed. Assessment no. 2 gives the target data value from WoS, which is the full publication name *Deutsche Medizinische Wochenschrift*, and assessment no. 3 gives the ISO abbreviated version of the publication name *Dtsch Med Wochenschr* (cf. section 5.3) decoded as *AbbrPubName_WoS* (AbbrPubName = abbreviated Publication Name; _WoS = from the WoS target record). The fifth row indicates the

[26] For reasons of clarity, the full author name is given in all examples. The actual data analysis distinguished between the fields last name, first initial and second initial.

assessment results, i.e. the inaccuracy codes (IACs) obtained by assessing the source data value *v* (i.e. Assessment nos. 1-3: *Deutscher Medizinwochenschrift*) against the target data values *v'* (i.e. Assessment no. 1: *Dtsch med Wschr*, assessment no. 2: *Deutsche Medizinische Wochenschrift* and assessment no. 3: *Dtsch Med Wochenschr*). The last row in the table is called *Explanation* which gives individual explanations to clarify the specific examples.

Table 9: General example table containing three assessment results

PubName_Ref	**Deutscher Medizinwochenschrift**		
Assessment no.	1	2	3
Data field	PubName_Orig	PubName_WoS	AbbrPubName_WoS
Data value	**Dtsch med Wschr**	**Deutsche Medizinische Wochenschrift**	**Dtsch Med Wochenschr**
Assessment result	I	B K Y	I
Explanation	The only valid result from comparing the value of the reference to the original value was that the reference does not give the same abbreviated title as the original article and was, therefore, assessed as *I* for *Abbreviation*.	A comparison of these two values allowed the detection of a *Spelling error* (IAC *B*: Deutscher vs. Deutsche); furthermore, Medizinwochenschrift and Medizinische Wochenschrift contain the same *Word stem* and were identified as the same journal. Hence, the assessment was IAC *Y* (*Word stem*). There is also a space character missing between Medizin and wochenschrift which was marked as IAC *K Space*.	The only valid conclusion from comparing these values was that the reference does not give the same abbreviated title as the WoS target record and therefore was assessed as *I* for *Abbreviation*.

The IAC *A* refers to a *Typographical variation*. *Typographical variation*s detected in this dataset include: spelling alternatives in American and British English (e.g.: behavior vs. behaviour or analyze vs. analyse, or Fungaemia vs. Fungemia), Latin vs. German spelling variations of medical terms (e.g. Metacarpophalangealgelenkes vs. Metakarpophalangeal-gelenkes) and spelling variations from other languages (e.g. Cusco vs. Cuzco). The inaccuracy category is not limited to the variations detected, but could be expanded in future work. The IAC *A* assesses string data values. It occurs as an assessment result in the following data field: article title.

The IAC *B* refers to a *Spelling error* that does not exceed the manipulation of two characters (on one string) in order to convert it into the correct value *v'*. Table 10 illustrates an example of an inaccurate article title where both assessments (against original and WoS target records) resulted in the IAC *B*. Another example is the following article title: *MACHIAVELLI AGAINST REPUBLICANISM. On the Cambridge Schools Guicciardinian Moments* vs. *Machiavelli against Republicanism. On the Cambridge Schools Guicciardinian Momemt*, where the second article title comes from a reference and actually contains two *Spelling error*s in one string (Momemt instead of Moments). Yet, the data manipulation does not exceed the exchange of two characters. In contrast, assessment no. 6 in Table 11 shows an article title that was assessed as IAC *J* (*Partially incorrect*) because the number of character edits exceeds two. The IAC *B* assesses string data values. It occurs as an assessment result in the following data fields: author name, article title, publication name.

Table 10: Example of IAC *B Spelling error*

ArticleTit_Ref	**Knowledge atitudes and practices of business travelers regarding malaria risk and prevention**	
Assessment no.	4	5
Data field	ArticleTit_Orig	ArticleTit_WoS
Data value	**Knowledge attitudes and practices of business travelers regarding malaria risk and prevention**	**Knowledge attitude and practices of business travelers regarding malaria risk and prevention**
Assessment result	B	B
Explanation	The string *atitudes* in the article title of the reference contains one *Spelling error* when assessed against the original article title. One character edit (adding the missing t) is necessary to translate the value into the correct value. Hence, a *B* was assigned.	The string *atitudes* in the article title of the reference contains two *Spelling error*s when compared to the WoS target value. Two character edits (adding the missing t and removing the s) are necessary to translate the value into the correct value. Hence, a *B* was assigned.

The IAC *C* stands for *Different language* which means that the two assessed data values are not in the same language. It was not further analyzed whether the source data value is a correct translation of the target data value. However, machine translation could be employed in the assessment process for this purpose. In assessment no. 7 in Table 11 the German article title from the source article is assessed against the English article title from the WoS target record. The IAC *C* for *Different language* was assigned. The IAC *C* assesses string data values. It occurs as an assessment result in the following data field: article title.

Table 11: Example of IAC *C Different language,* IAC *J Partially incorrect*

ArticleTit_Ref	**Industriesoziologie als Wirklichkeitssoziologie**	
Assessment no.	6	7
Data field	ArticleTit_Orig	ArticleTit_WoS
Data value	**Industriesoziologie als Wirklichkeitswissenschaft**	**Industrial sociology as science of reality**
Assessment result	J	C
Explanation	It takes more than two character edits to transform Wirklichkeitssoziologie into Wirklichkeitswissenschaft, therefore this value was assessed as *J*.	The IAC *C* stands for a *Different language* and, therefore, no further assessment could be carried out.

The IAC *D* denotes *Completely incorrect*. Any data value that was present, but where the original data value was no longer recognizable, was assessed as IAC *D*. *Completely incorrect* data values represent cases of semantic inaccuracy (cf. section 3.2). Table 12 gives an example in which a string data value is *Completely incorrect*; Table 13 illustrates an example in which a numerical data value is *Completely incorrect*. The IAC *D* assesses string data and numerical values. It occurs as an assessment result in all data fields.

Table 12: Example of IAC *D Completely incorrect* – string value

PubName_Ref	**Studies in Higher Education**		
Assessment no.	8	9	10
Data field	PubName_Orig	PubName_WoS	AbbrPubName_WoS
Data value	**Journal of Curriculum Studies**	**Journal of Curriculum Studies**	**J Curric Stud**
Assessment result	D	D	D
Explanation	The source data value does not match the target data value except for the word “studies”, therefore, the assessment resulted in *D* (*Completely incorrect*).	The source data value does not match the target data value except for the word “studies”, therefore, the assessment resulted in *D* (*Completely incorrect*).	The source data value does not match the target data value. It can be assumed that “Stud” could stand for “studies”, otherwise the values do not match at all. Therefore, the assessment resulted in *D* (*Completely incorrect*).

Table 13: Example of IAC *D Completely incorrect* – numerical value

EPage_Ref	**16**	
Assessment no.	11	12
Data field	EPage_Orig	EPage_WoS
Data value	**538**	**538**
Assessment result	D	D
Explanation	The data value from the reference is *Completely incorrect* and, therefore, the assessment result was *D*.	The data value from the reference is *Completely incorrect* and, therefore, the assessment result was *D*.

The IAC *E* refers to a completely *Omitted* data value. Note that a missing target data value is not assessed as IAC *E*, but as *Z Not available*, since the rule of the assessment direction dictates that source data values must be assessed against target data values. The IAC *E* assesses string data and numerical fields. It occurs as an assessment result in all data fields, except publication name.

The IAC *F* stands for *Cropped* which means that the data value is incomplete. Either the value is *Cropped* at the beginning, at the end, or both, but not in-between. The concept *Cropped* implies that the source data value as such is a correct part of the target data value, but one or more words are missing. In the case of author names, the IAC *F* indicates the use of *et al.*, i.e. the rest of the author names are not given due to the citation style in the reference. In article titles, the IAC *F* mostly stands for missing subtitles. Since the bibliographic fields in this data sample do not distinguish between main title and subtitle, an article title with a missing subtitle is assessed as *Cropped* (IAC *F*) and no separate inaccuracy category was established. Table 14 illustrates an example of a *Cropped* article title, Table 15 an example of a *Cropped* publication name, and Table 16 shows an example of a *Cropped* ending page. The IAC *F* assesses string data and numerical values. It occurs as an assessment result in the following data fields: author name, author first and second initial, article title, publication name, volume number, and ending page.

Table 14: Example of IAC *F Cropped* (article title), IAC *C Different language*

ArticleTit_Ref	**Reproduktion als Praxis**	
Assessment no.	13	14
Data field	ArticleTit_Orig	ArticleTit_WoS
Data value	**Reproduktion als Praxis Zum Vermittlungszusammenhang von Arbeits und Lebenskraft**	**Reproduction as individual action The interdependency of workstrength and lifestrength**
Assessment result	F	C
Explanation	The subtitle *Zum Vermittlungszusammenhang von Arbeits und Lebenskraft* is missing and, therefore, the assessment resulted in *F*.	The title is not in the same language and therefore was assessed as *C* for *Different language*.

Table 15: Example of IAC *F Cropped* (publication name)

PubName_Ref	**Heteroatom**		
Assessment no.	15	16	17
Data field	PubName_Orig	PubName_WoS	AbbrPubName_WoS
Data value	**Heteroatom Chemistry**	**Heteroatom Chemistry**	**Heteroatom Chem**
Assessment result	F	F	F
Explanation	Compared to the target data value, the publication name in the reference lacks the second part, *Chemistry*, and, therefore, an *F* for *Cropped* was assigned.	Compared to the data value from the WoS target record, the publication name in the reference lacks the second part, *Chemistry*, and, therefore, an *F* for *Cropped* was assigned.	Compared to the abbreviated publication name from the WoS target record, the publication name in the reference lacks the second part, *Chem*, and, therefore, an *F* for *Cropped* was assigned.

Table 16: Example of IAC *F Cropped* (ending page)

EPage_Ref	**43**	
Assessment no.	18	19
Data field	EPage_Orig	EPage_WoS
Data value	**543**	**543**
Assessment result	F	F
Explanation	In references, the format of pagination is often 540-43 or 540-3, which means that, compared to the original data value, the value from the reference is *Cropped*. Therefore, the IAC *F* for *Cropped* was assigned.	In references, the format of pagination is often 540-43 or 540-3, which means that, compared to the data value from the WoS target record, the value from the reference is *Cropped*. Therefore, the IAC *F* for *Cropped* was assigned.

The IAC *G* denotes *Interchanged fields*. This IAC is assigned whenever the target data values from different fields are exchanged within the same data record. One possible example is the interchange of any of the numerical fields, e.g. volume and issue numbers are switched or the starting page and ending page are interchanged. Since the issue number was not recorded from the original target record in the data collection process, the WoS target record of the cited article was consulted to verify whether a confusion with the issue number was the reason for the inaccurate source data values. Table 17 shows an example where the starting page is interchanged with the issue number. The IAC *G* is also assigned for switched first and second initials of the same author name. When the initials of two different authors from the same record are confused, a different IAC was applied (*O* for *Incorrect order of authors*). Table 18 illustrates an example of switched first and second initials. To keep track of which of the fields were interchanged, numbers were assigned additionally:

- IAC *G1*: this data field contains the issue number
- IAC *G2*: this data field contains the starting page
- IAC *G3*: this data field contains the ending page
- IAC *G4*: this data field contains the volume number
- IAC *G5*: this data field contains the last name
- IAC *G6*: this data field contains the first initial
- IAC *G7*: this data field contains the second initial

The IAC *G* assesses string data and numerical values. It occurs as an assessment result in the following data fields: author last name, first and second initial, volume number, starting and ending page.

Table 17: Example of IAC *G Interchanged fields* (starting page / issue number)

SPage_Ref	**2**	
Assessment no.	20	21
Data field	SPage_Orig	SPage_WoS
Data value	**189**	**189**
Assessment result	G1	G1
Explanation	The starting page given in the reference actually refers to the issue number of the cited article which was manually checked with the WoS target record as the issue number was not part of the data collection from the cited articles. Therefore, the IAC *G1* was assigned.	The starting page given in the reference actually refers to the issue number of the cited article. Therefore, the IAC *G1* was assigned.

Table 18: Example of IAC *G Interchanged fields* (first and second initial)

AuthorName_Ref	**Crans, CD**	
Assessment no.	22	23
Data field	AuthorName_Orig	AuthorName_WoS
Data value	**Crans, DC**	**Crans, DC**
Assessment result	G7 / G6	G7 / G6
Explanation	For the sake of clarity, the full author name is provided. Yet, the assessment result *G* was only assigned to the two data fields first (*G7*) and second initial (*G6*) and not to the author's last name.	For the sake of clarity, the full author name is provided. Yet, the assessment result *G* was only assigned to the two data fields first (*G7*) and second initial (*G6*) and not to the author's last name.

The IAC *H* decodes a *Jumbled* (data) *value*. The data field contains the correct value, but the order within the field is jumbled. Numerical data fields may hold transposed digits (e.g. starting page 564 vs. 654). In string data fields, the order of the strings is jumbled (e.g. main and subtitle: Die molekulare Welt des Lebensmittelgenusses. Auf den Geschmack gekommen vs. Auf den Geschmack gekommen. Die molekulare Welt des Lebensmittelgenusses), but each string value within the field is correct. The transposition of letters in a string falls into the category IAC *B Spelling error*. The IAC *H* assesses string data and numerical values. It occurs as an assessment result in the following data fields: author name, article title, publication name as well as starting and ending page.

The IAC *I* stands for *Abbreviation*, i.e. for any kind of abbreviation contained in a data field. It was mainly assigned to publication names. Many references in the citing articles use different *Abbreviation*s of the publication name, which do not necessarily correspond to the ISO standard abbreviation for journal titles. A data value was assessed as IAC *I* whenever the source data value contained an abbreviated version of the publication name and the target value gave the full name (assessment nos. 27 and 28, Table 20) or vice versa (assessment nos. 24 and 26, Table 19). If the *Abbreviation*s of the publication names from the target and source data values did not match, clearly referred to the same journal but did not conform to the ISO abbreviation, the IAC *I* was still assigned (assessment no. 29, Table 20). An example of a publication name corresponding to the ISO abbreviation of the journal name can be found in assessment no. 32, Table 21. The IAC *I* was also assigned to article titles that contained *Abbreviation*s when compared to the original article title (e.g. *Deutschsprachige Fassung und Validierung der Edinburgh postnatal depression scale* vs. *Deutschsprachige Fassung und Validierung der EPDS*). A possible further development of this IAC category could include chemical elements (in the pilot study (Olensky, 2013) article titles contained either the full name of a chemical element or the abbreviation from the periodic table). The IAC *I* assesses

string data values. It occurs as an assessment result in the data fields: article title and publication name.

Table 19: Example of IAC *I Abbreviation* – full publication name in source data value

PubName_Ref	**Chemie in unserer Zeit**		
Assessment no.	24	25	26
Data field	PubName_Orig	PubName_WoS	AbbrPubName_WoS
Data value	**Chem Unserer Zeit**	**Chemie in unserer Zeit**	**Chem Unserer Zeit**
Assessment result	I	0	I
Explanation	The cited article only provides the abbreviated publication name and as the reference in the citing article contained the full publication name, the IAC *I* for *Abbreviation* was assigned.	The data value is correct and, therefore, the assessment result was 0.	Compared to the ISO abbreviated publication name from WoS, the IAC *I* was assigned for the same reason as in assessment no. 24.

Table 20: Example of IAC *I Abbreviation* – abbreviated publication name in source data value

PubName_Ref	**Soc Inquiry**		
Assessment no.	27	28	29
Data field	PubName_Orig	PubName_WoS	AbbrPubName_WoS
Data value	**Sociological Inquiry**	**Sociological Inquiry**	**Sociol Inq**
Assessment result	I	I	I
Explanation	The reference from the citing article only provides an abbreviated publication name. In view of the cited article containing the full publication name, the IAC *I* for *Abbreviation* was assigned.	The reference from the citing article only provides an abbreviated publication name. In view of the data field from the WoS target record containing the full publication name, the IAC *I* for *Abbreviation* was assigned.	Even though the data values could be identified as a variation of the abbreviated publication name of the same publication, the source data value does not conform to the ISO abbreviation and, therefore, the IAC *I* for *Abbreviation* was assigned.

Table 21: Example of IAC *I Abbreviation* – ISO abbreviated publication name in source data value

PubName_Ref	**J Travel Med**		
Assessment no.	30	31	32
Data field	PubName_Orig	PubName_WoS	AbbrPubName_WoS
Data value	**Journal of Travel Medicine**	**Journal of Travel Medicine**	**J Travel Med**
Assessment result	I	I	0
Explanation	The reference from the citing article only provides an abbreviated publication name. In view of the cited article containing the full publication name, the IAC *I* for *Abbreviation* was assigned.	The reference from the citing article only provides an abbreviated publication name. In view of the data field from the WoS target record containing the full publication name, the IAC *I* for *Abbreviation* was assigned.	The reference contains the correct ISO abbreviation of the publication name and, therefore, does not contain any inaccuracies.

The IAC *J* denotes *Partially incorrect* data values. If a data value does not qualify as a *Spelling error B* (data manipulation threshold of two character edits), but the correct value is still recognizable, the data value was assessed as IAC *J Partially incorrect*. This definition excludes any stop words (cf. IAC *X Stop word*) as well as strings that have the same *Word stem* (cf. IAC *Y*). In contrast to the IAC *F* for *Cropped*, the IAC *J* for *Partially incorrect* decodes article titles in which parts of the article title were missing in-between, but not at the beginning or end of the article title. Assessment no. 6 in Table 11 shows an example where only a part of one string within the source value is inaccurate. Assessment no. 33 in Table 22 shows an example where two strings are missing in the source value. The IAC *J* assesses string data values. It occurs as an assessment result in the data fields: article title and publication name.

Table 22: Example of IAC *J Partially incorrect* (article title), IAC *B Spelling error*

ArticleTit_Ref	**Metaphysics in the Dark A Response to Rorty and Laclau**	
Assessment no.	33	34
Data field	ArticleTit_Orig	ArticleTit_WoS
Data value	**Metaphysics in the Dark A Response to Richard Rorty and Ernesto Laclau**	**Metaphysics in the dark A response to Richard Rorty and Ernesto Ladau**
Assessment result	J	B J
Explanation	The target value lacks two strings, namely the first names Richard and Ernesto, therefore the assessment result was *J* for *Partially incorrect*.	Compared to the WoS target record, the source data value lacks two strings, namely the first names Richard and Ernesto, therefore the assessment result was *J* for *Partially incorrect*. Additionally, the name Laclau contains a *Spelling error* which was assessed as IAC *B*.

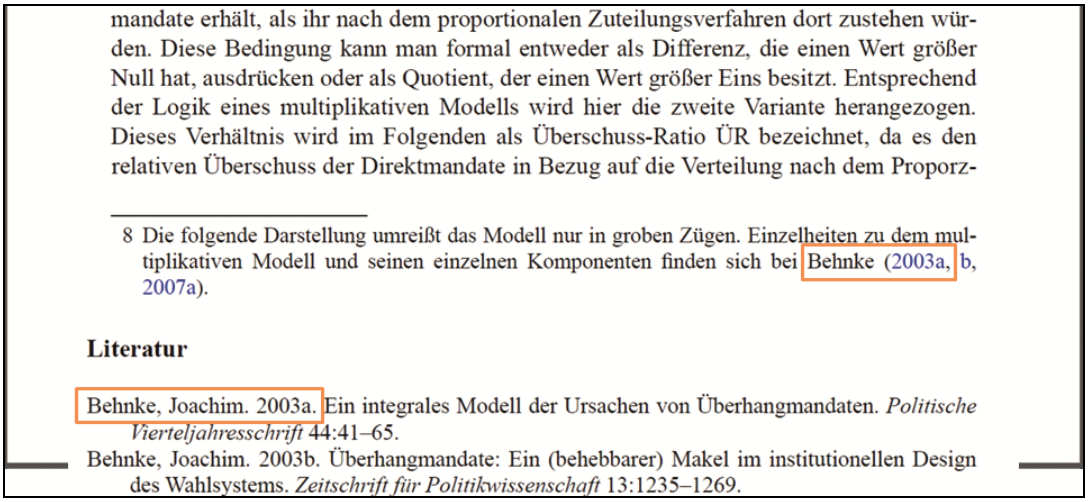
The IAC *K* refers to a *Space* character discrepancy. It is a minor inaccuracy that can easily be handled in the data parsing process. Again, multiple missing or additional *Space* characters were not counted as multiple inaccuracies. Assessment nos. 35 and 36 in Table 23 illustrate examples. The IAC *K* assesses string data values. It occurs as an assessment result in the data fields: author name, article title and publication name.

Table 23: Example for IAC *K Space* (article title)

ArticleTit_Ref	**Imported schistosomiasis in Europe sentinel surveillance data from Trop Net Europ**	
Assessment no.	35	36
Data field	ArticleTit_Orig	ArticleTit_WoS
Data value	**Imported Schistosomiasis in Europe Sentinel Surveillance Data from TropNetEurop**	**Imported schistosomiasis in Europe Sentinel surveillance data from TropNetEurop**
Assessment result	K	K
Explanation	The string “TropNet Europ” is spelled with additional space characters in the reference. Therefore, the data field was assessed as IAC *K Space*.	The string “TropNet Europ” is spelled with additional space characters in the reference. Therefore, the data field was assessed as IAC *K Space*.

The IAC *L* stands for *Informational letter*. It refers to letters that serve an informational purpose in numerical data fields in citations. They are part of a citation style mainly used in the SSH. Different publications by the same author in the same publication year are marked with additional letters (e.g.: 2003a, 2003b, 2003c, etc.), which facilitates a reference to the correct publication within the text. Figure 13 illustrates an example. References can also contain additional letters in their pagination (e.g. 61ff.). The IAC *L* assesses numerical data

values. It occurs as an assessment result in the data fields: publication year as well as starting and ending page.



mandate erhält, als ihr nach dem proportionalen Zuteilungsverfahren dort zustehen würden. Diese Bedingung kann man formal entweder als Differenz, die einen Wert größer Null hat, ausdrücken oder als Quotient, der einen Wert größer Eins besitzt. Entsprechend der Logik eines multiplikativen Modells wird hier die zweite Variante herangezogen. Dieses Verhältnis wird im Folgenden als Überschuss-Ratio ÜR bezeichnet, da es den relativen Überschuss der Direktmandate in Bezug auf die Verteilung nach dem Proporz-

8 Die folgende Darstellung umreißt das Modell nur in groben Zügen. Einzelheiten zu dem multiplikativen Modell und seinen einzelnen Komponenten finden sich bei Behnke (2003a, b, 2007a).

Literatur

Behnke, Joachim. 2003a. Ein integrales Modell der Ursachen von Überhangmandaten. *Politische Vierteljahresschrift* 44:41–65.

Behnke, Joachim. 2003b. Überhangmandate: Ein (behebbarer) Makel im institutionellen Design des Wahlsystems. *Zeitschrift für Politikwissenschaft* 13:1235–1269.

Figure 13: Example of IAC *L Informational letter*.

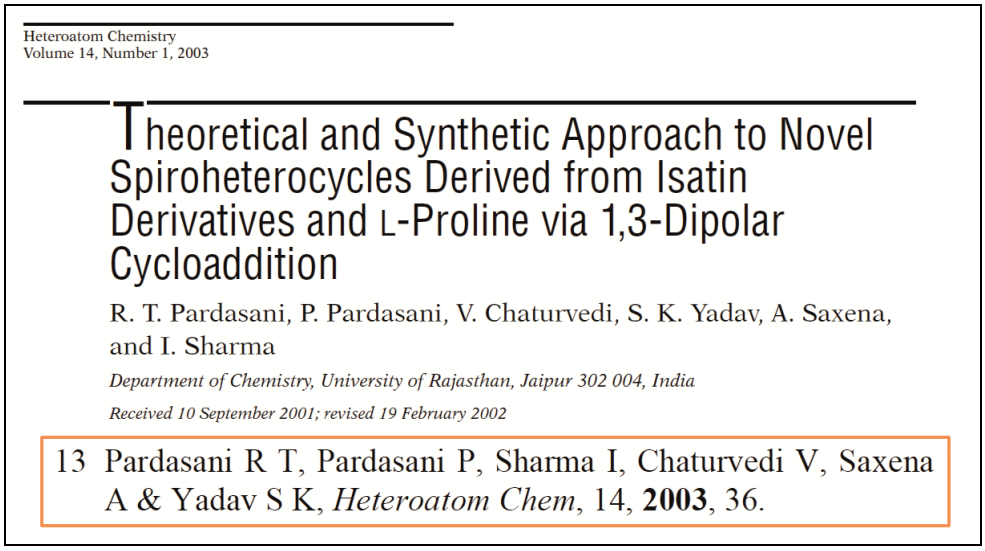
The IAC *M* denotes an *Incorrect interpretation of author names*. This could refer to interpreting parts of the last name as the middle name, i.e. second initial, or to interpreting the first name as the last name and vice versa. It mostly refers to last names consisting of multiple parts that are then interpreted incorrectly. In contrast to the IAC *G*, which marks *Interchanged fields*, the field values marked as IAC *M* cannot simply be switched. Assessment no. 37 in Table 24 gives an example of an incorrect interpretation of one part of the author's last name. In addition, this example illustrates how different the assessment results can be, depending on the target data values, since the assessment based on the WoS target record (assessment no. 38) does not show a discrepancy between the source and the target data value. Another example involves the incorrect interpretation of name prefixes such as *von*, *van* or *de* as first or middle names. The IAC *M* assesses string data values. It occurs as an assessment result in the data fields: authors' last name as well as first and second initial.

Table 24: Example of IAC *M Incorrect interpretation of author names* (first and second initial).

AuthorName_Ref	**Malar, EJP**	
Assessment no.	37	38
Data field	AuthorName_Orig	AuthorName_WoS
Data value	**Padma Malar, EJ**	**Malar, EJP**
Assessment result	M	0
Explanation	For the sake of clarity, the full author name is provided. Yet, the assessment result *M* was only assigned to the two data fields: last name and second initial.	For the sake of clarity, the full author name is provided. Comparing the reference to the values from the WoS target record the values are correct and, therefore, the assessment result for all three data fields (last name, first initial and second initial) was 0.

The IAC *N* refers to *Additional information* that is correct as such and in many cases useful for the reader, but is not actually part of the correct data value. Examples include: *in press*, *Review*, *in German*, *Suppl*, *Forthcoming* or a translation of the publication name (e.g. *Zeitschrift für Pädagogik Journal of Pedagogy*). Another example is the abbreviation *OPED* which was found as part of the data field article title. *OPED* stands for *opposite editorial page*, hence it provides *Additional information* for the reader, but is not part of the correct title. Often this *Additional information* can be detected during the data parsing process by looking for strings in round or square brackets. The IAC *N* assesses string data values. It occurs in the data fields: author second initial, article title and publication name.

The IAC *O* stands for an *Incorrect order of authors*. This assessment code was applied if the data values of different authors were switched. It could apply to all three of the author name fields or just to one of them. As mentioned in the description of IAC *G*, if the initials of one author were jumbled, the assessment result was *G* for *Interchanged fields* instead of *O* for *Incorrect order of authors*. Figure 14 illustrates an example of an *Incorrect order of authors*, where the authors in positions 3, 4 and 6 were jumbled. Analogously to the IAC *D Completely incorrect*, the IAC *O* is another example of a semantic inaccuracy. The IAC *O* assesses string data values. It occurs as an assessment result in the data fields: author name, first and second initial.



Heteroatom Chemistry
Volume 14, Number 1, 2003

Theoretical and Synthetic Approach to Novel Spiroheterocycles Derived from Isatin Derivatives and L-Proline via 1,3-Dipolar Cycloaddition

R. T. Pardasani, P. Pardasani, V. Chaturvedi, S. K. Yadav, A. Saxena, and I. Sharma

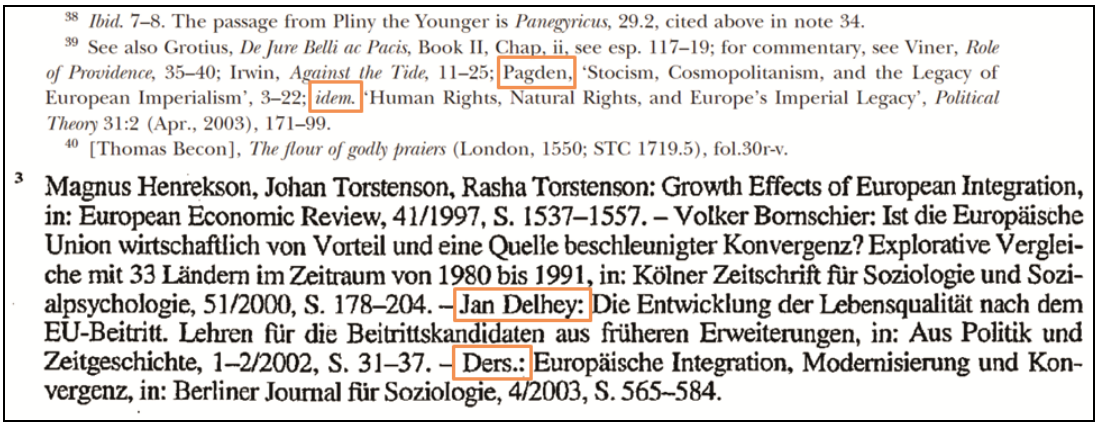
Department of Chemistry, University of Rajasthan, Jaipur 302 004, India

Received 10 September 2001; revised 19 February 2002

13 Pardasani R T, Pardasani P, Sharma I, Chaturvedi V, Saxena A & Yadav S K, *Heteroatom Chem*, 14, **2003**, 36.

Figure 14: Example of IAC *O Incorrect order of authors*

The IAC *P* decodes a *No author name* inaccuracy. It describes a phenomenon mainly known in the SSH where one author is cited in two subsequent citations with two or more different publications. Since the author name has already been given in the preceding citation, the author name is not repeated in the succeeding one. Different citation styles express the repetition of the preceding author name: e.g. ders. [in German], idem., ---------, ibid. Figure 15 illustrates two examples of authors who were cited in footnotes, each with two different publications in subsequent references. The IAC *P* assesses string data values. It occurs as an assessment result in the data fields: author name, first and second initial.



[38] *Ibid.* 7–8. The passage from Pliny the Younger is *Panegyricus*, 29.2, cited above in note 34.

[39] See also Grotius, *De Jure Belli ac Pacis*, Book II, Chap. ii, see esp. 117–19; for commentary, see Viner, *Role of Providence*, 35–40; Irwin, *Against the Tide*, 11–25; Pagden, 'Stocism, Cosmopolitanism, and the Legacy of European Imperialism', 3–22; *idem.* 'Human Rights, Natural Rights, and Europe's Imperial Legacy', *Political Theory* 31:2 (Apr., 2003), 171–99.

[40] [Thomas Becon], *The flour of godly praiers* (London, 1550; STC 1719.5), fol.30r-v.

3 Magnus Henrekson, Johan Torstenson, Rasha Torstenson: Growth Effects of European Integration, in: European Economic Review, 41/1997, S. 1537–1557. – Volker Bornschier: Ist die Europäische Union wirtschaftlich von Vorteil und eine Quelle beschleunigter Konvergenz? Explorative Vergleiche mit 33 Ländern im Zeitraum von 1980 bis 1991, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 51/2000, S. 178–204. – Jan Delhey: Die Entwicklung der Lebensqualität nach dem EU-Beitritt. Lehren für die Beitrittskandidaten aus früheren Erweiterungen, in: Aus Politik und Zeitgeschichte, 1–2/2002, S. 31–37. – Ders.: Europäische Integration, Modernisierung und Konvergenz, in: Berliner Journal für Soziologie, 4/2003, S. 565–584.

Figure 15: Example of IAC *P No author name*

The IAC *Q* stands for the assessment of *Special character*s and their spelling variations. *Special character*s detected in this dataset include: Germanic umlauts: Köster vs Koster,

Gjørup vs. Gjorup; other *Special characters*[27] from different alphabets like the German *ß* vs. *ss* or the Slavic *ğ*, and Roman numerals. Yet, this inaccuracy category is not limited to these characters and could be further expanded in future work. In contrast to the IAC *A Typographical variation*, which refers to linguistic spelling variations, the IAC *Q* denotes the technical representation of variants for special characters. Table 25 gives an example of a volume number in Roman numerals. The IAC *Q* assesses string data and numerical values. It occurs as an assessment result in the data fields: author name, second initial, article title, publication name, and volume number.

Table 25: Example of IAC *Q Special character* (Roman Numerals)

Vol_Ref	**XXXI**	
Assessment no.	39	40
Data field	Vol_Orig	Vol_WoS
Data value	**31**	**31**
Assessment result	Q	Q
Explanation	The volume number given in the reference is the Roman numeral equivalent to the Arabic number 31, which is the volume number given in the cited article. Therefore, the IAC *Q* for *Special character* was assigned.	The volume number given in the reference is the Roman numeral equivalent to the Arabic number 31, which is the volume number given in the WoS target record. Therefore, the IAC *Q* for *Special character* was assigned.

Table 26: Example of IAC *S Padded* (article title), IAC *C Different language*

ArticleTit_Ref	**Das Berner Konzept Die Reorientierung der dysplastischen Hüftpfanne durch die *Berner periazetabuläre Osteotomie* nach Ganz**	
Assessment no.	41	42
Data field	ArticleTit_Orig	ArticleTit_WoS
Data value	**Berner periazetabuläre Osteotomie**	**The Bernese periacetabular osteotomy**
Assessment result	S	C
Explanation	For the sake of clarity, the correct value that is contained in the reference of the citing article was italicized. This *Padded* version of the article title was therefore assessed as IAC *S*.	The title is not in the same language and therefore a *C* was assigned which stands for *Different language*.

The IAC *R Punctuation* denotes differing punctuation used in a data value. It was only assigned in the CitedRef assessment processes, since, for the other processes, all punctuation

[27] For a list of non-identifiable special characters by the Levenshtein distance function, cf. Table 40, in Appendix D.

was eliminated in the data parsing process. The IAC *R* assesses string data values. It occurs as an assessment result in the data fields: author names and first initial.

The IAC *S* decodes a *Padded* data value. The source data field contains the correct value, but also contains additional values that are identified as incorrect and are of no additional value to the reader (in contrast to IAC *N Additional information*). In most cases, the origin of these added data values is not reproducible. For instance, if the correct ending page number is 193 and the corresponding target value is 5193, the assessment result is the IAC *S*. In contrast to the IAC *J Partially incorrect*, a data value that was assessed as *S* contains the complete correct value plus additional incorrect strings or numbers. The IAC *J* on the other hand does not contain the complete correct value. Assessment no. 41 in Table 26 gives an example of a *Padded* article title. The IAC *S* assesses string data and numerical values. It occurs as an assessment result in the data fields: second initial, article title, publication name, volume number as well as starting and ending page.

The IAC *T Plus/Minus* denotes a data value that is correct if the number 1 or 2 is added to, or subtracted from, the data value. The calculation can either be made on the total number or just on one of the digits. In contrast to transposed digits (covered by IAC *H*), the data manipulation includes a mathematical operation instead of just switching two digits. For example, if the target data value of the ending page number is *169* and the source data value is *167*. By adding 2 to *167* one gets *169*, which coincides with the target data value. Therefore, this data value was assessed as *T* for *Plus/Minus*. Moreover, if the reference in the value is *269*, the calculation can be performed on the first digit only and will result in the source value (subtracting 1 from the first digit: 2 minus 1 = 1; which converts the number *269* into *169*). This inaccuracy pattern also occurs in publication years: e.g. *1998* vs. *1988* or *1998* vs. *1999*. In both cases the correct publication year could be calculated. The IAC *T* assesses numerical values. It occurs as an assessment result in the data fields: publication year, volume number as well as starting and ending page.

The IAC *U Full first name* denotes a value in the first initial field where the full first author name is given instead of the initial. This was again a special case in the assessment of the cited reference information (CitedRef). The IAC *U* assesses string data values. It occurs as an assessment result in the data field first initial.

The IAC *V* stands for an *Incorrect interpretation of additional information*. It is related to the IAC *M Incorrect interpretation of author names* and to the IAC *N Additional Information.* In

contrast to the IAC *M*, the IAC *V* decodes any kind of information that was interpreted incorrectly. The IAC *M* only designates author-related information interpreted incorrectly. In contrast to the IAC *N Additional Information*, the IAC *V* refers to data values in data fields that were interpreted as (part of) valid values for one of the assessed bibliographic fields, but have no informational value for the reader. They may originate from an automatic data collection process. Figure 16 provides an example in which the affiliation "Stephen F. Austin State University" of the actual author "Jerry Williams" is separated by a comma from the author name, which can be seen in the lower part of the figure. The comma separation usually indicates the next author. Apparently, the automatic data extraction process interpreted the affiliation as an additional author, which is shown screenshot from WoS in the background. Therefore, this author name was assessed as IAC *V*. The IAC *V Incorrect interpretation of additional information* was also assigned if additional information from the authors was interpreted as part of the article title: *Travelers Knowledge Attitudes and Practices on Prevention of Infectious Diseases Results from a Pilot Study* vs. ***European Travel Health Advisory Board*** *Travelers knowledge attitudes and practices on prevention of infectious diseases results from a pilot study*. The IAC *V* assesses string data values. It occurs as an assessment result in the data fields: author name, first and second initial and article title.

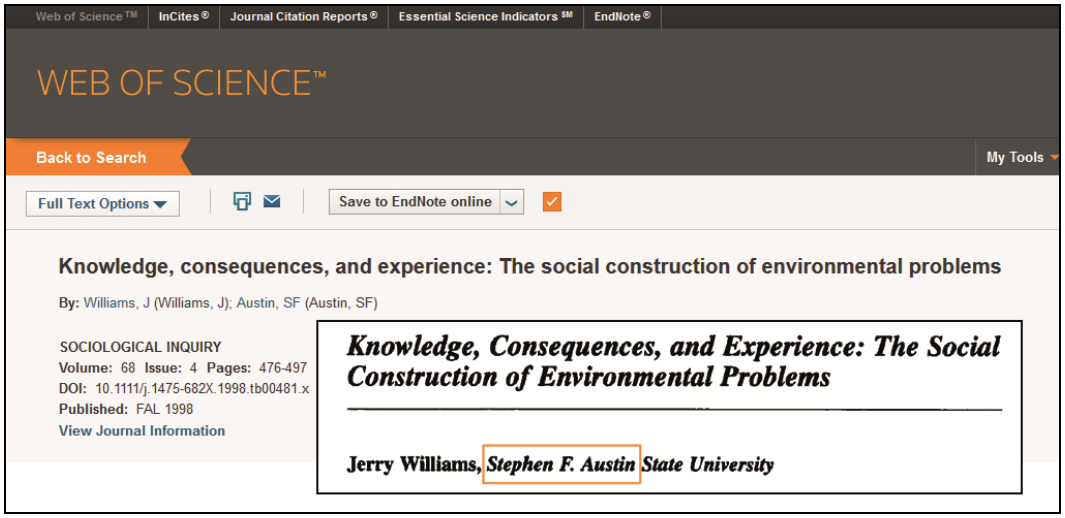


Figure 16: Example of IAC *V Incorrect interpretation of additional information*

The IAC *X Stop word* decodes small inaccuracies like the omission, addition or jumbling as well as spelling mistakes in, and of, stop words (e.g. in, on, of, the, and[28]) in string data values.

[28] A list of all stop words that were identified in the assessment process can be found in Appendix D.

It occurs as an assessment result in the data fields: author last name (e.g. Fachkommission Diabetes in Sachsen vs. Fachkommission Diabetes Sachsen), article title and publication name.

The IAC *Y Word stem* is located between the IAC *B Spelling error* and the IAC *J Partially incorrect*. It decodes assessments where the *Word stem* of a string is the same but the ending is not. Furthermore, it exceeds the two-character edit limit (which is the criterion for inaccuracies to be decoded as *B Spelling error*). Both Table 9, assessment no. 2 and Table 27 show examples of *Word stem* inaccuracies. The IAC *Y* assesses string data values. It occurs as an assessment result in the data fields: article title and publication name.

Table 27: Example of IAC *Y Word stem* (article title)

ArticleTit_Ref	**Its your most precious thing worstcase thinking trust and parental decision making about vaccines**	
Assessment no.	43	44
Data field	ArticleTit_Orig	ArticleTit_WoS
Data value	**Its Your Most Precious Thing WorstCase Thinking Trust and Parental Decision Making about Vaccinations**	**Its Your Most Precious Thing Worstcase thinking trust and parental decision making about vaccinations**
Assessment result	Y	Y
Explanation	The words vaccines and vaccinations are not the same data values but have the same *Word stem*. Therefore, the assessment result was *Y Word stem*.	The words vaccines and vaccinations are not the same data values but have the same *Word stem*. Therefore, the assessment result was *Y Word stem*.

The IAC *Z* decodes an assessment result that is *Not available*. It assesses missing data values in the target data fields. Assessment no. 46 in Table 28 gives an example of a *Z* assessment. The IAC *Z* assesses string data and numerical values. It occurs as an assessment result in the data fields: volume number and ending page.

Table 28: Example of IAC *Z Not available*

EPage_Ref	**238**	
Assessment no.	45	46
Data field	EPage_Orig	EPage_WoS
Data value	**238**	+
Assessment result	0	Z
Explanation	The data value is correct and therefore the assessment result was 0.	The ending page in WoS is missing and the field only contained the value +. Therefore, it is not possible to assess the value in the reference and the assessment result was IAC *Z* (*Not available*).

6.3 Taxonomy of inaccuracies in bibliographic references

The final codebook consists of 25 main IACs that were grouped into a taxonomy, which reflects a conceptual hierarchy. In contrast to a decision tree, the different groups are not valid values of the analysis (Krippendorff, 2004), but only serve to build a conceptual organization. First, the IACs were grouped according to the required level of sophistication of data manipulation rules, i.e. how elaborate data manipulation or the matching logic must be in order to match data values correctly. This impact was determined by whether the assessed data field actually contained a correct value (*simple*), contained part of a correct value (*moderate*) or did not contain a correct value (*complex*). So far, the categorization in reference error studies of minor, intermediate and major errors has been based on how severe the author perceived the error to be or if it impeded immediate retrieval (e.g. Goldberg et al., 1993; Oermann, Cummings & Wilmes, 2001). This taxonomy, on the other hand, is based on the degree of data accuracy and inaccuracy respectively.

The IACs in the three main groups (or categories) were clustered according to specific characteristics of the inaccuracy (e.g. disarranged data values or spelling variations), which constitute the middle-level categories. We will refer to them as subgroups or subcategories. Figure 17 illustrates the resulting taxonomy of bibliographic inaccuracies.

Starting in the first main group (*simple*), *Added data values* summarize all IACs that mark data values where values were added to the correct data value: *K Space*, *L Informational letter*, *N Additional information*, *R Punctuation*, *U Full first name* and *S Padded*. *Disarranged data values* include all IACs where data values were, in principle, correct, but their order had been jumbled within the data fields: *G Interchanged fields*, *H Jumbled value* and *O Incorrect order of authors*. Since the data fields of both subcategories contain the correct values, with the right data parsing rules, such as excluding space characters from the matching procedure or excluding letters from numeric fields (e.g. publication year), it should be possible to identify and match the correct value.

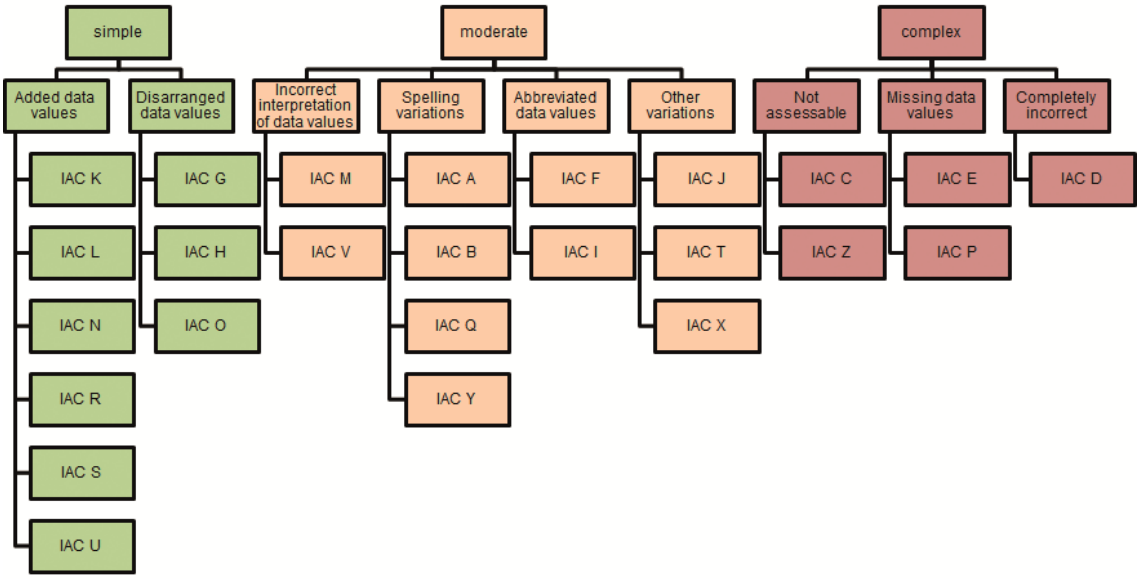


Figure 17: Taxonomy of bibliographic inaccuracies

Moving on to the second main group (*moderate*), *Incorrect interpretation of data values* summarizes the IACs *M Incorrect interpretation of author names* and *V Incorrect interpretation of additional information*. In contrast to *Disarranged data values*, these data values cannot be corrected by simply switching fields or the order within a field, but require more sophisticated data manipulation. For example, if part of an author's last name is interpreted as the given name, only the initial will be saved in the data record. Hence, it is more difficult to find, identify and correct these inaccuracies. *Spelling variations* are grouped together in another subcategory, covering: *A Typographical variation*, *B Spelling error*, *Q Special character* and *Y Word stem*. All of these variations entail data manipulation rules that would need to employ additional resources like dictionaries or thesauri. The subcategory *Abbreviated data values*, summarizes the IACs *F Cropped* and *I Abbreviation*. They both mark data values that have not been fully spelled out and require other resources (e.g. list of ISO abbreviations of journal names) in the data manipulation process. *Other variations* include data values that also entail more sophisticated data manipulation rules in order to convert an inaccurate value into a correct one or to match it. The IACs in this subgroup are: *J Partially incorrect*, *T Plus/Minus*, *X Stop word*. All subcategories could also employ mechanisms able to recognize that the present data values are part of the correct value (e.g. calculation of *n*-grams or longest common substring, cf. section 2.2.3).

In the third main group (*complex*), the subgroup *Not assessable* summarizes IACs that mark data values as impossible to further assess and, therefore, match. This includes data values that are in a *Different language* (IAC *C*) or *Not available* (IAC *Z*) in the target record. Even though elaborate tools, such as automatic language detection and machine translation, could make data values marked as IAC *C* assessable, these methods seem to bear no relation to the expected outcome, especially if we consider that the IAC *C* only occurs in article titles, which are seldom used in the citation matching process. The IAC *E Omitted* (IAC *E*), denoting any missing data values, and the IAC *P*, denoting an author name replaced by a place holder specific to a citation style, are grouped together as *Missing data values*. Last, the IAC *D Completely incorrect* forms its own subgroup, as no other inaccuracies are similar to this specific case. In all three subgroups, it will be hard to find suitable data manipulation rules in order to use the data values for citation matching.

6.4 Summary

This chapter set out to answer the research question how data accuracy in a bibliometric data source can be assessed and how the identified bibliographic inaccuracies can be categorized. To this end, we combined the automatic assessment process with a qualitative content analysis (Parts A and B), explained in section 5.2. The coding procedure defines a set of measures that were taken in order to guarantee validity, objectivity and reliability of the analysis. In three assessment iterations (Steps 1 and 2 plus one iteration from Step 3), the codebook was established, pinpointing the inaccuracies found in the citing references. The final assessment iteration of Step 3 assigned the IACs to the data values. The organization of the IACs into a taxonomy of three main groups and nine subgroups further summarizes the inaccuracies according to common attributes, such as *Spelling variations* or *Missing data values*. The taxonomy is the basis for the quantitative analysis of inaccuracies in chapter 7 and chapter 8. It also impacts the proposals on how citation matching processes could be improved (cf. chapter 9).

7 Quantitative Analysis of Bibliographic Inaccuracies

This chapter discusses the quantitative results of the qualitative content analysis. Section 7.1 reports on the results of the evaluation of the WoS against the original target records (Orig-WoS). Section 7.2 summarizes the occurrences of inaccuracies in the two assessment samples, where the source records were assessed against both target datasets (Orig-Ref and WoS-Ref). The frequencies of all IACs are discussed in detail and reasons for differences in the results are given. Furthermore, the inaccuracies occurring in the different bibliographic fields are explained. In the following sections 7.3-7.8 the results are analyzed according to the different strata of the data sample: *domain*, *discipline* and *language of the cited article* as well as *document type*, *language* and *citation window of the citing article*. They highlight differences and distinctions within the facets (e.g. NS vs. SSH in the facet *domain*). The inaccuracies are analyzed according to the subcategories from the taxonomy presented in section 6.3: *Added data values*, *Disarranged data values*, *Incorrect interpretation of data values*, *Spelling variations*, *Abbreviated data values*, *Other variations*, *Not assessable*, *Missing data values* and *Completely incorrect*. Detailed result tables are given in Appendix F. Section 7.9 discusses the consolidation of the two assessment variants into one result set. The Orig-Ref sample contained a variant of the bibliographic field *article title* for some records; the WoS-Ref sample contained two variants for some records: one of the bibliographic field *article title* and one of the bibliographic field *publication name* (cf. section 5.3 for a detailed description of the variants). Section 7.10 discusses the occurrences of false positives. Section 7.11 summarizes the findings of this chapter.

Since the goal of the analysis was to fully grasp the structure of bibliographic data in references and the related inaccuracies, but not to determine an overall accuracy rate of the records or databases, we analyzed the occurrences of the IACs based on all data values and did not summarize them per record. The number of inaccuracies was normalized by the number of assessed data values present in each evaluated instance of a facet. For example, in the facet

document type, we found 6,139 inaccuracies distributed over 42,075 data values in *Articles* and 90 inaccuracies distributed over 534 data values in *Book Chapter / Book* (figures are from the Orig-Ref result set). To compare the shares of inaccuracy subcategories by source data values, the number of inaccuracies in each subcategory (e.g. *Added data values*, *Disarranged data values*, etc.) was normalized by the number of IACs present in each category, as shown in Table 29.

Table 29: Number of IACs per inaccuracy category

Inaccuracy category	Number of IACs
simple	
Added data values	4 (6)[29]
Disarranged data values	3
moderate	
Incorrect interpretation of data values	2
Spelling variations	4
Abbreviated data values	2
Other variations	3
complex	
Not assessable	2
Missing data values	2
Completely incorrect	1

During the analysis, a few irregularities[30] were discovered in the data: two out of the 300 cited articles as well as 20 citing articles had an incorrect article language assigned in WoS. As the language of the cited and citing articles was documented during the data entry process, we could verify the language provided by WoS. The articles were categorized according to the corrected languages for the evaluation in sections 7.5 and 7.7.

7.1 Evaluation of original article vs. WoS record

This section discusses the results of the assessment of the WoS target records against the original target records (Orig-WoS) of the 300 cited articles. Table 42 in Appendix F contains the results of the overall frequency of IACs in the Orig-WoS result set. 90% of all WoS data values do not contain a discrepancy, i.e. are completely accurate, when compared to the

[29] Since the IACs *R Punctuation* and *U Full first name* only occurred in the CitedRef-WoS and the CitedRef-Sco result, the number of IACs in the subcategory *Added data values* was 6 instead of 4.
[30] All irregular WoS records are documented in Appendix H. They were also reported to Thomson Reuters for correction.

bibliographic data of the original article. Logically, all inaccuracies are reflected in the assessment results of the Orig-Ref and WoS-Ref datasets. Depending on the reference, the assessment result may conform to the data given in the original article or in the WoS record.

The WoS records contain completely accurate publication years, volume numbers and starting pages. Inaccuracies found in the ending pages are primarily related to missing data values (IAC *E Omitted*) and also to a few transposed values (IAC *T Plus/Minus*). The missing ending page number resulted in *Not assessable* data values in the WoS-Ref result (IAC *Z Not available*). Inaccuracies in the publication name are caused by abbreviations different from those used in the original article (IAC *I Abbreviation* and IAC *X Stop word*). That way, we discovered that the abbreviations of two German journals in the original article did not correspond to the correct ISO abbreviation. The predominant discrepancy, however, in publication names as well as in authors' last names is the IAC *Q Special character*, which is caused by the different handling of German umlauts in WoS. A few *Special character*s also occur in article titles. The prevailing IAC in article titles is *C Different language*, which denotes the article titles from the German dataset that had been translated into English in WoS. Hence, in the WoS-Ref result the correct German article title in a reference, if not translated, resulted in *C Different language*. One article title is *Cropped* (IAC *F*), one contains *Additional information* (IAC *N*) and two contain a *Spelling error* (IAC *B*). However, the majority of *Spelling error*s occurs in authors' last names. Other IACs that occur in small quantities in the authors' names are the following: the IAC *M Incorrect interpretation of author names* (2 WoS records), where compounded names were interpreted differently; IAC *O Incorrect order of authors* (1 WoS record); IAC *V Incorrect interpretation of additional information* (1 WoS record), where *OPED* was interpreted as part of the article title when it actually stands for *opposite the editorial page*; and one case in which a last and a first name were switched and another where a first and a second initial were switched (IAC *G Interchanged fields*). Additionally, a few records lack the first or second initials of author names (IAC *E Omitted*) or the first or second initials contain *Completely incorrect* data values (IAC *D*). The IACs *A Typographical variation*, *H Jumbled value*, *J Partially incorrect*, *K Space*, *L Informational letter*, *P No author name*, *Y Word stem* do not occur at all.

To summarize, all IACs from the Orig-WoS result were reflected in the result sets Orig-Ref and WoS-Ref. In the WoS-Ref result, correct data values in the references were assessed as inaccurate because of incorrect WoS data values. Incorrect data values in the references, if conforming to the incorrect WoS data values, resulted in a correct assessment in the WoS-Ref result, but multiplied the IAC from the Orig-WoS in the Orig-Ref result. Therefore, the data

from the two sets of target articles provides differing results for the assessment of the references. These differences are discussed in the following section 7.2.

7.2 Evaluation of overall occurrences of IACs

This section describes the overall distribution of inaccuracies in the two result sets (Orig-Ref and WoS-Ref). The total number of citing references assessed (including the missed citations identified in the *Cited Reference Search*), i.e. source records, is 3,929 which come from 3,735 different citing articles[31]. In the Orig-Ref result, we found slightly fewer inaccuracies than in the WoS-Ref sample, while the number of assessed data values is approximately the same. 85% of all data values do not contain a discrepancy in either dataset, yet only 18% (Orig-Ref) and 15% (WoS-Ref) of all source records are discrepancy-free. In the WoS-Ref sample, the number of records with one or two inaccuracies accounts for more than half of the records, whereas in the Orig-Ref sample the share is below 50%. We investigated the occurrences of IACs within the subcategories defined in the taxonomy of bibliographic inaccuracies (cf. section 7.2.1) as well as their occurrences in the different bibliographic fields (cf. section 7.2.2).

7.2.1 Discussion of IACs

This subsection ranks the inaccuracy subcategories according to the size of their shares within each assessment result and discusses differences between the two result sets Orig-Ref and WoS-Ref (Figure 18). Within each subcategory, the absolute numbers of IAC occurrences (cf. Table 43, Appendix F) are additionally compared, as they reveal more information about the structure of data in WoS, the original articles and the references.

Inaccuracies are more evenly distributed in the WoS-Ref result set than in the Orig-Ref result set, which shows two distinctive peaks in the *Abbreviated data values* and the *Missing data values* subcategories. We found most inaccuracies in the category *complex* for the WoS-Ref result (54%), but in the category *moderate* for the Orig-Ref result (57%). The group *moderate* contains 39% of all inaccuracies in the WoS-Ref result; whereas *complex* inaccuracies account

[31] 91% of the citing articles contained only one reference to one of the 300 investigated cited articles, 7% of the citing articles contained two relevant references, 1% contained three references and there were 12 articles with four references to cited articles as well as five articles with five references to cited articles.

for 39% in the Orig-Ref result. The fewest inaccuracies were detected in the category *simple* (Orig-Ref: 4%; WoS-Ref: 7%).

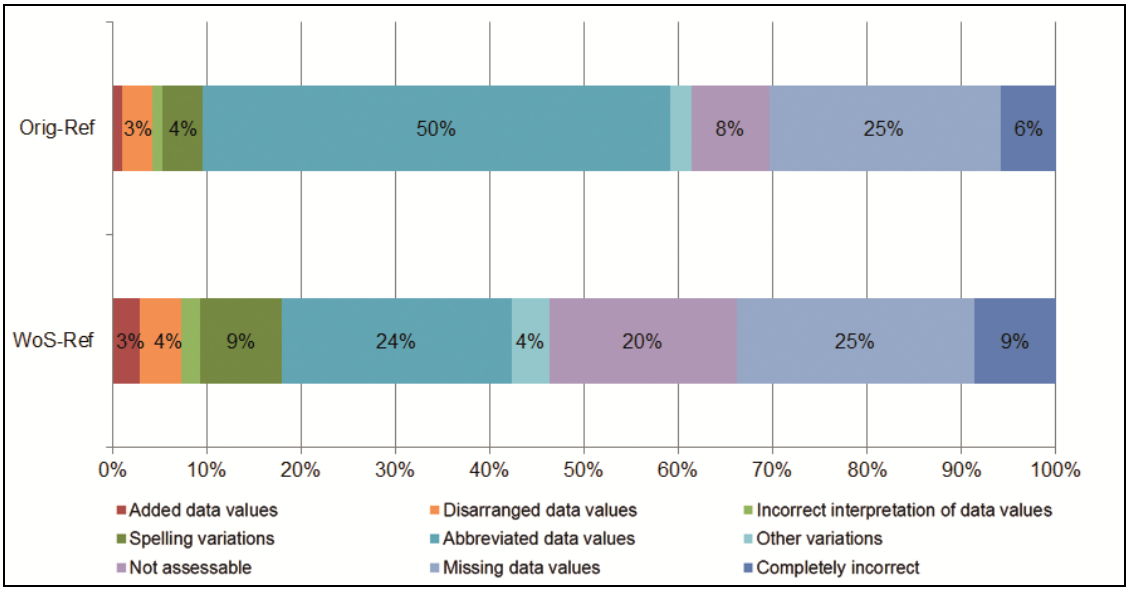


Figure 18: Overall shares of inaccuracy subcategories (source data value level)[32]

Abbreviated data values are the most frequent in the Orig-Ref result set, followed by *Missing data values*. In the WoS-Ref result the shares of these two categories is almost the same. Within the subcategory *Abbreviated data values* the distribution between the two IACs *F Cropped* and *I Abbreviation* is reversed for the two result sets: the IAC *F* is more frequent in the WoS-Ref result, whereas in the Orig-Ref result the IAC *I* is more frequent. Both IACs show a decrease in absolute numbers in the WoS-Ref result compared to the Orig-Ref result. The occurrences of the IAC *F Cropped* decrease less drastically and can be explained by references that translated German article titles into English and could only be fully assessed in the WoS-Ref assessment process. The large decrease in occurrences of the IAC *I Abbreviation* is due to the fact that, in the WoS-Ref assessment process, the publication name was assessed against the full publication name and the ISO abbreviated version and the most accurate result was recorded (cf. 7.9.2 Evaluation of publication names and their abbreviations). In contrast, the correctness of abbreviated publication names could not be checked in the Orig-Ref result. Hence, they were assessed with the IAC *I Abbreviation*. The *Missing data values* subcategory is dominated by the IAC *E Omitted* in both assessment results. The IAC *P No author name* only occurs in 8 records and traces back to a specific citation style for citing a previously mentioned author. Some of the author-related *Missing data values* stem from incorrectly added

[32] Data labels of percentages are displayed in the figures if the shares are 3% or higher.

author names in the WoS target record. The absolute numbers for the IACs *E Omitted* and *P No author name* are (almost) the same in both result sets.

The third most frequent subcategory in both result sets is *Not assessable*. Within this, the IAC *C Different language* is the most frequent one. It shows an increase in occurrences from the Orig-Ref to the WoS-Ref result set, which means that more citing references use the original German article title than a translated English version. The IAC *Z Not available* occurs more often in the WoS-Ref result than in the Orig-Ref result because it stands for different target data values that were *Not available* for the assessment. As described in section 7.1, in the WoS target records they represent a missing ending page; in the Orig-Ref result they stand for volume numbers not printed in the original article title.

The subcategory *Completely incorrect*, consisting only of one IAC (*D Completely incorrect*), comes fourth in the ranking in the Orig-Ref result, whereas in the WoS-Ref result, this category ties with *Spelling variations* at 9% each. The share and the absolute number of the IAC *D* are higher in the WoS-Ref result, which can mainly be attributed to German articles that were cited with *Completely incorrect* English translations of their titles and *Omitted* second initials in the WoS target records. Therefore, second initials, that are in fact correct, were assessed as *Completely incorrect*.

In the Orig-Ref result, the fifth most frequent subcategory is *Spelling variations*. The most common IAC in this category is *Q Special character*, followed by *B Spelling error* in both result sets. In the Orig-Ref result *Typographical variation*s (IAC *A*) rank third, followed by *Y Word stem*. In the WoS-Ref result the ranking of these two IACs is reversed. Fairly small differences can be found between the absolute numbers of the two result sets of the IACs *A Typographical variation* and *B Spelling error*s. More interesting are the IACs *Q Special character* and *Y Word stem*, where the increase in the absolute numbers of the two result sets is more than 100% from the Orig-Ref to the WoS-Ref sample. Since WoS does not handle special characters consistently (e.g. the German umlaut *ä* is transliterated either as *a* or *ae*), the number of IAC *Q* is higher in the WoS-Ref than in the Orig-Ref result, which also means that references tend to use the accurate spelling of special characters as given in the original article. The marked difference in the occurrences of IAC *Y Word stem* can be explained by references where German article titles had been translated into English. Hence, they could not be further assessed in the Orig-Ref assessment process (IAC *C Different language*).

Next in the ranking of the subcategories comes *Disarranged data values* in the Orig-Ref result, whereas in the WoS-Ref result this category ties again with another subcategory, *Other variations*, at 4% each. The IAC *O Incorrect order of authors* occurs the most often, followed by *G Interchanged fields* and *H Jumbled value* in both assessment results. The two result sets provide similar absolute numbers for the IACs *G Interchanged fields* and *H Jumbled value*. The *Jumbled value*s primarily trace back to one author name occurring in two cited articles with the German prefix *von*. The author was cited as *Ferber von, L* instead of *von Ferber, L*. However, this different way of citing the name may have been induced by the required citation style. On looking further into the detailed results of the *Interchanged fields* (IAC *G*), a similar distribution is revealed in both sets: in half of the cases, the inaccuracies are interchanges of author-related fields (Orig-Ref: 50%; WoS-Ref: 45%), whereas issue numbers that are mistaken for the volume number or the starting page are a little less frequent in the Orig-Ref result (43%) than in the WoS-Ref result (48%). The starting page number, mistakenly entered as the volume number or the ending page number, accounts for 5% in the Orig-Ref and 4% in the WoS-Ref sample. The ending page number occurring in the field of the starting page number is the least frequent (2% for both sets). The highest difference in absolute occurrences is found in the IAC *O Incorrect order of authors*, where the result of the WoS-Ref sample accounts for more occurrences than that of the Orig-Ref sample. This divergence can be explained by one cited article, where the author order is disarranged in the WoS target record as a consequence of interpreting the first author as two separate ones[33]. Hence, references citing the correct author order were assessed as discrepant, even though the order was actually correct.

In the Orig-Ref result, the seventh most frequent subcategory is *Other variations*, in which the IAC *T Plus/Minus* occurs the most, followed by the IAC *J Partially incorrect*. In the WoS-Ref result set the ranking of these two IACs is reversed. The IAC *X Stop word* occurs the least in both assessment results. The IAC *T Plus/Minus* shows a small decrease in absolute numbers in the WoS-Ref result compared to the Orig-Ref one, caused by *Omitted* ending page numbers in the WoS target record, which impeded any further assessment of some references. In contrast, the absolute occurrences of the other two IACs increase significantly in the WoS-Ref result set. Analogously to the IAC *Y Word stem*, both increases result from references where German article titles had been translated into English which were assessed as IAC *C Different language* in the Orig-Ref assessment process, but could be assessed in more detail in the WoS-Ref assessment process.

[33] The first author named Ngoc Hoa Tran Huy was interpreted as Hoa, N. and Huy, T (cf. Appendix H, Table 82).

The last two subcategories rank differently in the two result sets: in the Orig-Ref result the share of *Incorrect interpretation of data values* is larger than that of *Added data values*. In the WoS-Ref set, the subcategories swap places. Within the *Incorrect interpretation of data values* the IAC *M Incorrect interpretation of author names* occurs more often than the IAC *V Incorrect interpretation of additional information*. The different absolute numbers of IAC *M Incorrect interpretation of author names* can again be explained by discrepancies in the WoS target records. In particular, parts of some authors' last and given names in five cited articles are interpreted differently than given in the original article (cf. Table 83, Appendix H). As the absolute number is almost doubled compared to the Orig-Ref result, more citations follow the interpretation of names in the original article than that used by WoS. In the two assessment samples, the IAC *V Incorrect interpretation of additional information* only occurs in article titles, where names of consortia were added to article titles when they should actually be part of the author information (two cited articles had a consortium in their author information). The slight difference in the absolute numbers of the two samples traces back to German article titles that could only be assessed as IAC *C Different language* in the WoS-Ref assessment. Hence, the IAC *V* assessment results from the Orig-WoS result did not multiply in the WoS-Ref result, but resulted in *Missing data values* for the additional author.

The IACs within the subcategory *Added data values* also rank differently for the two result sets, except for the IAC *K Space*, which occurs the least in both result sets. In the Orig-Ref result, the IAC *S Padded* is the most frequent IAC, followed by the IAC *L Informational letter* and *N Additional information*. In the WoS-Ref result set, the IAC *N Additional information* leads the ranking, followed by the other two IACs. However, no big differences in the absolute numbers of the two result sets can be observed for the two IACs *K Space* and *L Informational letter*. The IAC *K Space* predominantly occurs in article titles with a few exceptional occurrences in authors' last names. The IAC *L Informational letter* describes a specific citation style for publication years and starting page numbers. The discrepancy in the results for the IAC *N Additional information* (for the reader) is interesting. Compared to the Orig-Ref result, the significantly higher number of data values assessed with IAC *N* in the WoS-Ref set can be explained by one article by *Arduengo, AJ III* and *Krafczyk, R*. In the original article, the suffix *III* is counted as part of the field second initial, and many references cited the author correspondingly. Yet, the WoS target record does not contain the suffix and, therefore, these data values were assessed with the IAC *N Additional information*[34]. Other *Additional*

[34] It could be argued that, as the additional value consists of Roman numerals, it could be interpreted as IAC *Q Special character*. However, the IAC *Q* was only used when the source record also contained

information mainly refers to additions, such as *in press* or *in German* in the field publication name. The difference in the result for the IAC *S Padded* (increase of 87% in the WoS-Ref result over the Orig-Ref set) can be explained by assessments of non-translated English article titles (resulting in IAC *C Different language* in the Orig-Ref assessment) and abbreviated publication names (resulting in IAC *I Abbreviation* in the Orig-Ref assessment). The IACs *R Punctuation* and *U Full first name* only occur in the CitedRef result sets (cf. section 8.3).

7.2.2 Discussion of IACs in bibliographic fields

While section 7.2.1 has already related most of the inaccuracy subcategories to the bibliographic field in which they occur most frequently, this subsection describes the categories occurring in each bibliographic field (cf. Figure 19; Table 44 and Table 45 in Appendix F). The findings specifically influence the proposals in chapter 9, since they reveal the nature of inaccuracies that need to be dealt with in the citation matching in each bibliographic field. In particular, the results of the bibliographic fields not typically used in citation matching processes will shed light on what a change of policy could imply in terms of data matching rules due to additional inaccuracies.

In the first author fields, the most inaccuracies were found in *Missing data values* in the Orig-Ref result and in *Spelling variations* in the WoS-Ref result, but *Missing data values* rank second. *Missing data values* were mainly detected in first and second initials and a few author names were missing altogether (IAC *P No author name*). The difference in the share of *Spelling variations* again reflects the different handling of *Special character*s (IAC *Q*) in WoS. *Abbreviated data values* in the Orig-Ref result and *Added data values* in the WoS-Ref result represent the different ways of citing the suffix *3rd* to one author's last name (see previous section: *Arduengo* example). The *Incorrect interpretation of author names* has a larger share in the WoS-Ref than in the Orig-Ref set, which can be explained by some discrepant WoS target records compared to the original article. *Disarranged data values* as well as *Completely incorrect* data values have an almost equal share in both result sets. *Completely incorrect* values occur predominantly in second initials, followed by first initials. *Disarranged data values* denote author names which have been cited in an incorrect order (IAC *O*) or where first and second initial have been interchanged (IAC *G*).

---

the *Special character*, but with an orthographic variation, e.g. 3rd instead of III, which was only the case for the original source record, but not for the WoS target record.

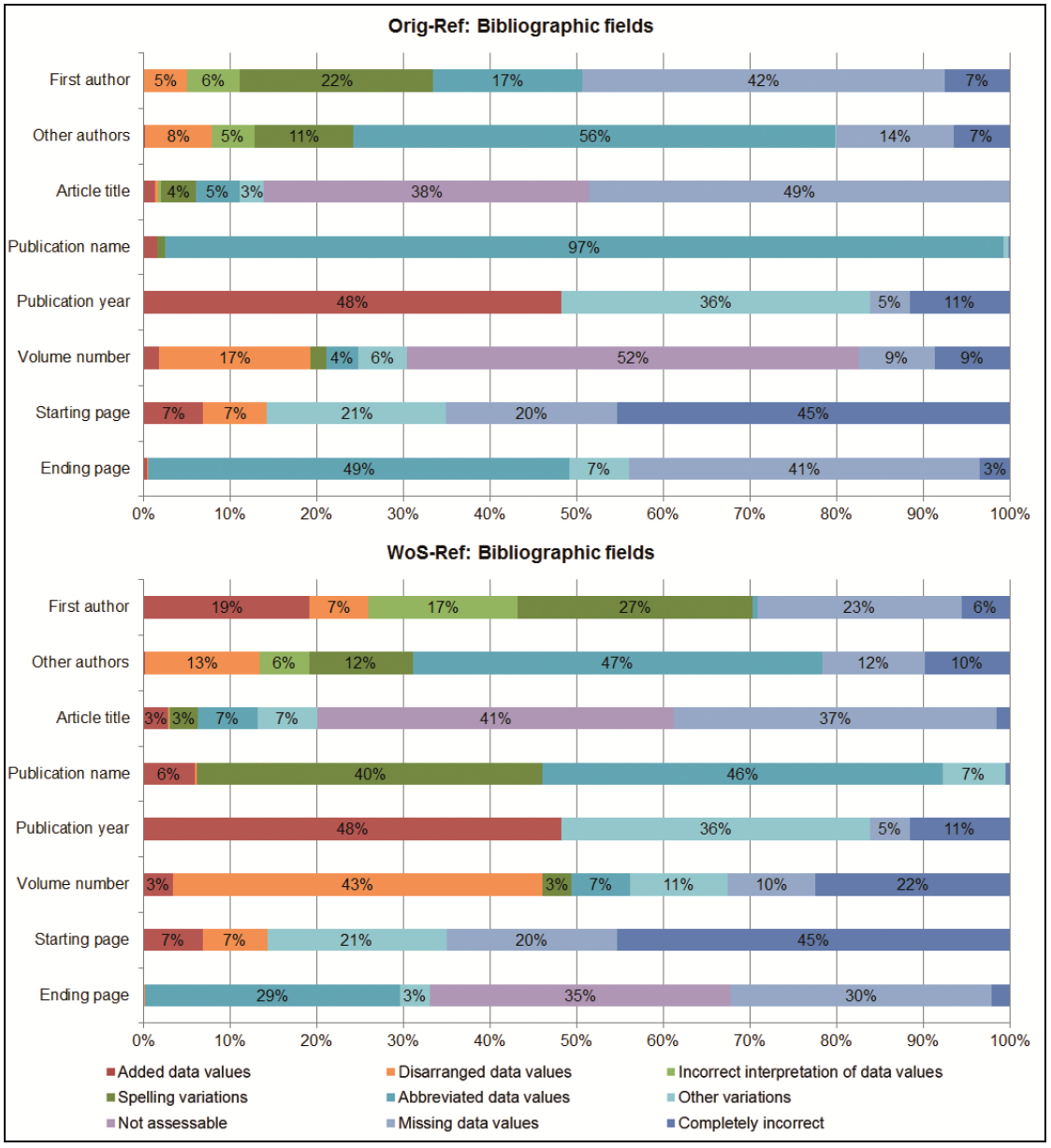


Figure 19: Shares of inaccuracy subcategories per bibliographic field (source data value level)

In the other author-related fields (authors 2-23), the distribution is similar. The high percentage of *Abbreviated data values* in both result sets can be attributed to the use of *et al.* in the bibliographic references. *Completely incorrect* occurs slightly more often in the WoS-Ref result due to second initials cited in the references which were not present in the WoS target records. Apart from that, *Missing data values*, *Spelling variations*, *Disarranged data values* and *Incorrect interpretation of author names* also occur and their causes do not deviate from those described for the first author fields.

The article title has large shares of inaccuracies assessed as *Not assessable* and *Missing data values*. *Missing data values* are caused by the citation style, dominating in the NS (cf. section 7.3) where it is common not to cite the article title at all. *Not assessable* can be ascribed to the language differences and translations of article titles in the German dataset (IAC *C Different language*). Other inaccuracies reflect the characteristics of an article title consisting of a long string: *Abbreviated data values* (omission of subtitle), *Other variations* and *Spelling variations* (caused by *Spelling error*s, differences in American and British English spellings, special characters and *Partially incorrect* parts of the title) and *Added data values* (imaginative citing authors adding non-existent subtitles). As mentioned in section 7.2.1, the small percentage of *Completely incorrect* data values traces back to German articles which were cited with *Completely incorrect* English translations of their titles and could only be assessed in the WoS-Ref assessment process.

The most inaccuracies in publication names are caused by the different *Abbreviation*s used in bibliographic references (*Abbreviated data values*), specifically in the Orig-Ref result. However, in the WoS-Ref result, these inaccuracies also amount to almost half of all inaccuracies, which means that a considerable number of references uses a different abbreviation of the publication name than the ISO one. *Other variations* refer to discrepancies due to *Stop word* (IAC *X*), where the values may also be a type of *Abbreviation*. *Added data values* can be explained by added information for the reader (IAC *N*), such as *in German* or *in press*. Again, *Spelling variations* have a larger share in the WoS-Ref result than in the Orig-Ref result, as they refer to German umlauts occurring in German publication names (e.g. Berliner Journal f**ü**r Soziologie). The publication name is *Completely incorrect* in only one reference.

For the publication year, the shares of all inaccuracy subcategories are the same in both result sets. *Added data values* have the largest share, which is explained by the IAC *L Informational letter* and originates from the citation style. *Other variations* are related to publication years deviating by one or two years from the original publication year (IAC *T Plus/Minus*). 0.25% of publication years are cited *Completely incorrect* and in even fewer records, which are related to references to forthcoming publications, the year is *Omitted* all together (IAC *E*).

Since volume numbers are missing in some of the original articles (IAC *Z Not available*), the shares of inaccuracies according to the WoS-Ref data represent a more accurate assessment result and, therefore, are discussed here. The largest share of inaccuracies in volume numbers is attributed to *Disarranged data values*, stemming mainly from issue numbers that were

mistaken for the volume number. The second biggest problem is *Completely incorrect* volume numbers, followed by *Other variations*, i.e. numbers that could be calculated correctly (IAC *T*) and *Missing data values*, which are not only related to references to forthcoming publications, but are also simply omitted from the reference. The fewest inaccuracies in volume numbers are attributed to *Abbreviated data values*, i.e. *Cropped* numbers (IAC *F*), *Spelling variations*, i.e. Roman numerals instead of Arabic numerals (IAC *Q Special character*) and *Added data values*, i.e. added values that are not part of the correct volume number and are also not related to any other numerical field (IAC *S Padded*).

For the starting page number, the shares of all inaccuracy subcategories are the same in both result sets. The most inaccuracies are caused by *Completely incorrect* data values, which implies that a data manipulation rule to convert incorrect into correct values will be almost impossible to find. However, in the evaluation of missed citations (cf. chapter 8) we discuss the share of cited page numbers in these *Completely incorrect* data values, which may provide a solution to parts of the incorrect values. An *Omitted* starting page (IAC *E*) or a starting page which only differs by one or two digits (*Other variations*, IAC *T Plus/Minus*) are other major issues. The remaining inaccuracies in starting page numbers are evenly distributed over *Disarranged* and *Added data values*. *Disarranged data values* mainly stem from issue numbers (and a few ending page numbers, i.e. IAC *G3*) mistakenly entered in the starting page number (IAC *G1*) as well as a few transposed digits (IAC *H Jumbled value*). *Added data values* originate, on the one hand, from the citation style as IAC *L Informational letter* (e.g. *pp.* 44-52); on the other hand, analogously to the volume number, they are related to randomly added values (IAC *S Padded*).

Since ending page numbers are missing in some of the WoS target records, the shares of inaccuracies of the Orig-Ref data represent a more accurate assessment result and, therefore, are discussed here. *Abbreviated data values* have the largest share of inaccuracies, which originate from the citation style only citing the last or last two digits of the ending page (e.g. 536-8 or 678-90), closely followed by *Omitted* ending pages. 7% of the inaccuracies are attributed to *Other variations*, i.e. the correct data value could be calculated (IAC *T Plus/Minus*). Very few ending page numbers are *Completely incorrect*.

To compare how strong the influence of inaccuracies in the different bibliographic fields on the records is, we evaluated how many records did not contain a discrepancy in the first author name (last name, first and second initial), all author names, article title, publication name, publication year, volume number and starting and ending page (cf. Table 30). 97% of the

source records contained a correct publication year and starting page (not necessarily in the same records). In 95% (Orig-Ref) and 97% (WoS-Ref) of all records the volume number is completely accurate. The difference between the two result sets is caused by *Not available* (IAC *Z*) volume numbers in original target articles in the Orig-Ref assessment sample. A high number of records also had a completely accurate first author (Orig-Ref: 90%; WoS-Ref: 86%), while all authors were absolutely correct for 74% of all records in the Orig-Ref result set and for 70% in the WoS-Ref result set. Around half of all records had no discrepancy in their article title. Even though this seems a low percentage in comparison to the other fields, it has to be borne in mind that the article title contains the longest data values and, therefore, the risk of containing inaccuracies is greater. The ending page is another bibliographic data field which is subject to an increased quantity of discrepancies, which can be partly explained by citation styles only requiring the last or last two digits of the ending page (e.g. 536-8 or 678-90). Only 43% of the records in the Orig-Ref result contained a completely accurate publication name, while the publication names in the WoS-Ref result with 76% discrepancy-free records in that field are more accurate. This difference can be ascribed to the fact that some references cite the full publication names and some the abbreviated ones. In the WoS-Ref result both variants could be assessed, whereas in the Orig-Ref result only one variant could be assessed, which, therefore, in many cases led to the IAC *I Abbreviation*.

Table 30: Share of 100% accurate bibliographic fields (source record level)

% of records without discrepancy		in the bibliographic field
Orig-Ref	**WoS-Ref**	
90%	86%	First author-related fields
74%	70%	All author-related fields
56%	46%	Article title
43%	76%	Publication name
97%	97%	Publication year
95%	97%	Volume number
97%	97%	Starting page
61%	51%	Ending page

A comparison of the bibliographic fields typically used in the citation matching process (first author name, publication name, publication year, volume number and starting page), reveals that not all IACs occur. The IACs *A Typographical variation*, *V Incorrect interpretation of additional information* and *C Different language* do not occur at all in bibliographic fields, since they are all specific to the article title field. Additionally, over 90% of occurrences of the IACs *F Cropped* and *Y Word stem* are attributed to fields not used in the citation matching

process. For the IAC *F Cropped*, this mainly refers to author-related fields (use of *et al*.) and the article title (omission of sub-title). The IAC *Y Word stem* primarily occurs in the article title. In contrast, IACs which occur only in fields that are used for citation matching are the IAC *Z Not available* in the Orig-Ref result set because it denotes missing volume numbers in the target articles, the IAC *I Abbreviation* in both results in the publication name and the IAC *L Informational letter* in both results in the publication year and starting page (even though there are a few additional occurrences in the ending page in the Orig-Ref result).

7.3 Evaluation per domain of the cited article

This section discusses the two result sets based on the *domain of the cited article*[35] (cf. Appendix F, Table 45-Table 47). 64% of all references cite an article related to the NS, while 36% of the references cite an article from the SSH, which corroborates the finding of other studies that the NS are better covered in WoS than the SSH (e.g. Norris & Oppenheim, 2007; Harzing, 2013a). References to SSH articles are more accurate than those to NS articles: in the Orig-Ref result set, 67% of all inaccuracies were detected in the NS and 33% in the SSH, whereas the difference between the domains in the WoS-Ref result is lower: NS: 53% and SSH: 47%. 43% (Orig-Ref) and 32% (WoS-Ref) of all source records in the SSH do not contain a discrepancy. Only 3% of the source records in the NS are discrepancy-free in the Orig-Ref result, and 6% in the WoS-Ref result. In the Orig-Ref result set, 54% of all source records in the SSH contain 1-3 inaccuracies per record, whereas this is true for 77% of the source records in the NS. Similarly in the WoS-Ref result set, 76% of all source records in the NS contain 1-3 inaccuracies, and this applies to 63% of all source records in the SSH.

Figure 20 illustrates the distribution of inaccuracies in the two domains NS and SSH for both result sets. *Abbreviated data values* occur more often in the NS than in the SSH in both datasets. In the Orig-Ref sample, this subcategory even constitutes 47% of all inaccuracies in the NS. This corroborates the observation that, in the NS, authors tend to use an abbreviated publication name in the reference, as well as crop the list of cited authors by using *et al*. *Not assessable* data values have a higher share in the NS than in the SSH in the Orig-Ref result set, which means that in contrast to references to SSH target articles, references to NS target records tend to contain translated foreign article titles. The WoS-Ref corroborates this with a

[35] As mentioned at the beginning of the chapter, for the evaluation of each facet, the number of inaccuracies was not only normalized by the number of IACs in each subcategory, but also by the number of assessed data values present in the evaluated instance of the facet.

reversed distribution of shares between NS and SSH. It also reflects missing ending page numbers in WoS. *Missing data values* are more frequent in the NS than in the SSH in both data samples, which can be explained by the prevailing citation style in the NS of not citing the article title and/or the ending page.

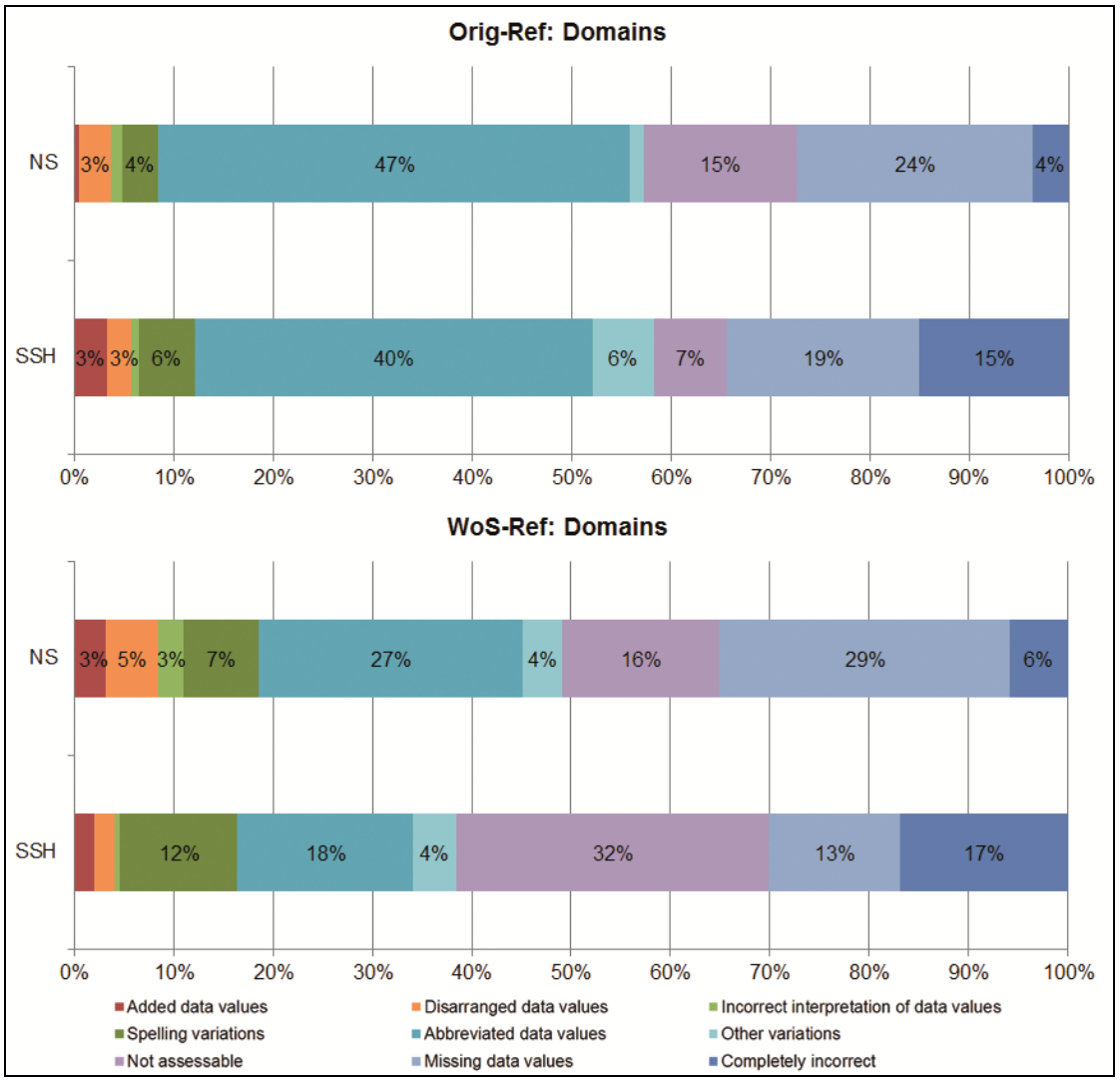


Figure 20: Inaccuracy subcategories per domain of the cited article (source data value level)

In both data samples, *Completely incorrect* data values, which are primarily found in starting and ending pages, have a larger share in the SSH than in the NS. *Other variations* are more frequent in the SSH than in the NS in the Orig-Ref sample, while in the WoS-Ref sample their shares are almost equal in both domains. Both results emphasize the commonly employed citation style in the SSH of merely providing the cited page number instead of the entire pagination. In the both results, the subcategory *Spelling variations* has a higher share in the SSH than in the NS, but the absolute numbers and shares are larger in the WoS-Ref result set.

Hence, SSH tend to cite *Special character*s correctly according to the original article. In the Orig-Ref result set, the SSH have a higher share of *Added data values* than the NS; the WoS-Ref result set presents a reversed picture. In the categories *Disarranged data values* and *Incorrect interpretation of data values*, both result sets provide similarly small shares, which occur more often in the NS than in the SSH.

The results reveal that citing authors in the SSH tend to follow the exact bibliographic data from the original article, whereas in the NS the WoS target records seem to be a more accurate match. Furthermore, they also reflect the different citation styles of the NS and SSH: in the NS the tendency is to shorten the bibliographic data, whereas in the SSH all the bibliographic data is reproduced more accurately according to the original bibliographic data. Although references to SSH target articles contain fewer inaccuracies, they have a higher share of *Completely incorrect* data values than the NS.

7.4 Evaluation per discipline of the cited article

This section summarizes the results of the quantitative analysis of inaccuracies according to the *discipline of the cited article* (cf. Appendix F, Table 48-Table 54). The different disciplines represented in the data sample are Multidisciplinary Chemistry (Chem), Orthopedics (Ortho), Internal & General Medicine (GenMed), Education & Educational Research (Edu), Political Science (PolSci) and Sociology (Sociol). Figure 21 depicts the shares of source records, i.e. citing references, which were assessed per discipline. As mentioned in the data selection process (section 5.5), the majority of references come from the NS disciplines, as they are better covered in the WoS than the SSH disciplines. However, the distribution is still fairly equal for all six disciplines, which reflects a balanced data sample and speaks for the comparability of the results.

On comparing the shares of inaccuracies (cf. Figure 32, Appendix F), the largest is found in Multidisciplinary Chemistry (Orig-Ref: 25%; WoS-Ref: 23%). In the Orig-Ref dataset, Internal & General Medicine and Orthopedics take second place with 20% each. The least inaccuracies are found in the SSH disciplines Sociology (14%), Political Science (12%) and Education & Educational Research (8%). The distribution of shares in the WoS-Ref dataset is different and less distinct. The second largest shares of inaccuracies are found in Political Science and Orthopedics (17%), followed by Internal & General Medicine tying with

Sociology at 15% each. The most accurate discipline is also Education & Educational Research with 13%.

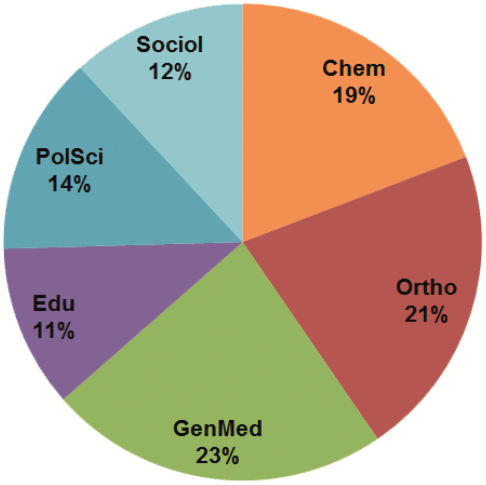


Figure 21: Shares of source records per discipline

Figure 22 summarizes the shares of inaccuracies per inaccuracy subcategory for each discipline. In general, the results of the evaluation per discipline reflect the overall differences between the Orig-Ref and WoS-Ref assessment process. If the results are the same in both assessment results, they characterize the citation style or specific inaccuracy patterns in each discipline. Differences in the results trace back to peculiarities in the target data values discussed in sections 7.1 and 7.2.

In Multidisciplinary Chemistry a rare use of article titles and citation of the ending page can be observed, which is reflected by a large share of *Missing data values* (IAC *E Omitted*) and a low share of *Not assessable* (*IAC C Different language*) in both result sets. The extensive use of *et al.* in the citations also contributes to the large share of *Missing data values*. The low share of *Abbreviated data values* in the WoS-Ref result reveals that, if a publication name is cited as an abbreviation, it tends to follow the ISO abbreviation. The highest share of *Incorrect interpretation of data values* is evidence that WoS target records assessed as IAC *M* or *V Incorrect interpretation of information* in the Orig-WoS assessment process originate from Multidisciplinary Chemistry and, therefore, passed the assessment result on to correct references. In Orthopedics, the relatively large share of *Spelling variations* in the WoS-Ref result is explained by the German dataset, which included a publication name with a German umlaut: Der Orthopäde. Since the share of *Abbreviated data values* in Orthopedics is larger than in Multidisciplinary Chemistry, we infer that citing authors of orthopedic articles do not

tend to use the full or the correct ISO abbreviated publication name as often. The large share of *Not assessable* data values shows that, in Orthopedics, article titles are cited in the reference more often than in Multidisciplinary Chemistry and, if so, they are primarily cited with their original article title and not translated. Additionally, the majority of non-assessable ending pages occurred in Orthopedics in the WoS target records. The rather low share of *Missing data values* indicates that references to orthopedic cited articles are in most cases complete and do not tend to omit bibliographic data. In Internal & General Medicine, *Abbreviated data values* have the largest share in the WoS-Ref result. This reflects, on the one hand, a higher tendency of citing authors to abbreviate the cited article titles compared to other disciplines. On the other hand, it suggests that citing authors use different abbreviations of publication names than the ISO one. In contrast to Orthopedics, citing authors tend to translate German article titles into English when citing them, which is expressed by the number of *Not assessable* data values. The proportionally large share of *Missing data values* can be traced to one cited article with 23 authors, which were not fully cited by all references and, therefore, resulted in a large number of *Missing data values*.

The largest share of *Spelling variations* in the WoS-Ref result is found in Education & Educational Research. Analogously to Orthopedics, this is explained by the German publication name containing two umlauts: Zeitschrift für Pädagogik. Education & Educational Research is the discipline with the most accurate citations of publication names and the least *Cropped* article titles (*Abbreviated data values*). The large share of *Other variations* stems from one cited article that was repeatedly cited with a transposed publication year as well as from references citing the cited page number instead of the entire pagination. The difference in the share of *Missing data values* for the two result sets reflects *Omitted* second initials in the references which were present in the original article, but not in the WoS target record. The large share of *Completely incorrect* data values in the WoS-Ref result is caused by one WoS target record with one *Completely incorrect* author name, which resulted in IAC *D Completely incorrect* for correct references. In Political Science, none of the German article titles were translated into English (cf. non-existent share of *Not assessable* in the Orig-Ref result). The rather high shares of *Missing data values* and *Completely incorrect* are a result of the frequent use of citing page numbers in references to cited articles in Political Science. *No author name* or other information was interpreted incorrectly in the Political Science references (*Incorrect interpretation of data values*). In addition to the fact that no compounded name occurred in Political Science, which is most probably a coincidence and not related to the discipline, it is the discipline with the second lowest number of author names per article. Hence, the probability to encounter a compounded name is lower than in other disciplines. Political

Science and Sociology are both SSH disciplines which tend to use different abbreviations for the publication names than the ISO abbreviation or the full publication names, as opposed to the other SSH discipline Education & Educational Research. The number of *Not assessable* data values in Sociology is not related to translated German article titles, but to original articles from which the volume number could not be extracted and was, therefore, not assessed in the references. The higher share of *Spelling variations* in the WoS-Ref result traces back to the German publication name containing one umlaut.

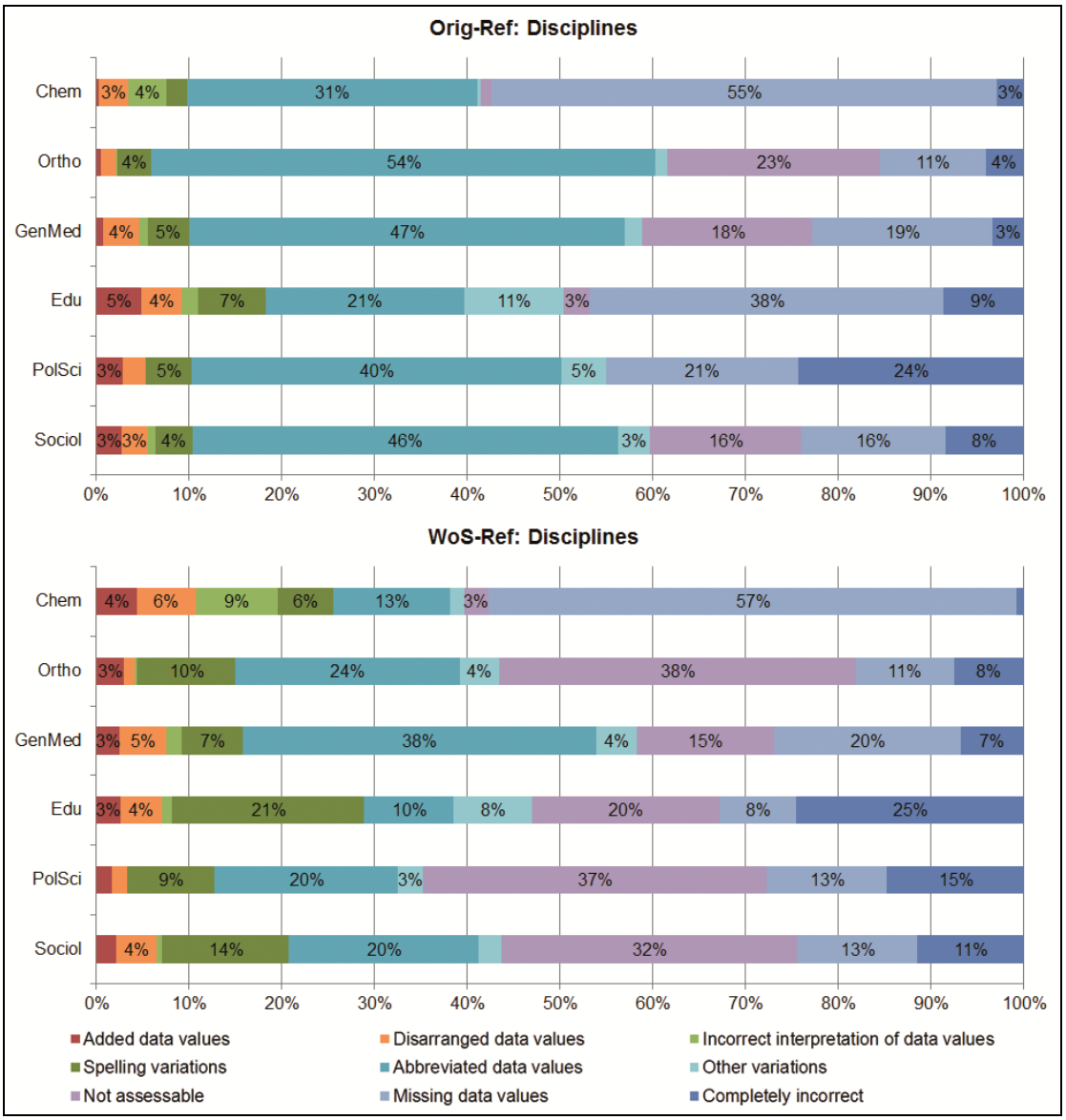


Figure 22: Inaccuracy subcategories per discipline of the cited article (source data values)

Disarranged data values have relatively even shares in all disciplines. Hence, they are not specifically influenced by the discipline or the citation style. In contrast, *Other variations* and

Completely incorrect data values occur more often in the SSH disciplines, which reflects the citation behavior of citing the cited page number rather than the entire pagination. While *Spelling variations* due to *Special character*s (IAC *Q*) are influenced by umlauts in the publication name of German-language journals and, therefore, multiply for each reference, other IACs in this category, such as IAC *A Typographical variation* or *B Spelling error* cannot be attributed to a specific discipline. In all three SSH disciplines, citing authors used the original article title in over 90% of the references, which is mirrored in the relatively large shares of *Not assessable* data values in the WoS-Ref result.

7.5 Evaluation per language of the cited article

This section documents the results of the evaluation based on the *language of the cited article* (Appendix F, Table 55-Table 57). 57% of all references cited an English article and 43% cited a German one. In both result sets, more than half of the inaccuracies are attributed to references to German cited articles, i.e. references to English cited articles are more accurate. References to German cited articles also have a lower share of records without any discrepancy in either result set.

Figure 23 summarizes the shares of inaccuracies per subcategory for the two languages of cited articles, English and German. The subcategories *Abbreviated data values* and *Disarranged data values* are slightly more frequent in references to English cited articles in the Orig-Ref result, whereas in the WoS-Ref result the difference between the two languages is more distinct. The *Incorrect interpretation of data values* category has a larger share in the English than in the German set which is caused by the IAC *M Incorrect interpretation of author names* in both result sets. This means that, in our data sample, compounded names which can lead to an incorrect interpretation, only occur in the references of the English subset. However, we doubt that this describes a general characteristic of references to English target articles. Additionally, the higher share of *Added data values* in the German subset of the WoS-Ref result only reflects one specific case of an author name discrepancy, described in section 7.2 as the *Arduengo* example, but not a pattern of references to German target articles containing more *Added data values*. The results of the *Not assessable* category are, on the one hand, due to the missing data values in the target records; on the other hand, they are related to German article titles being cited as translations or in the original language.

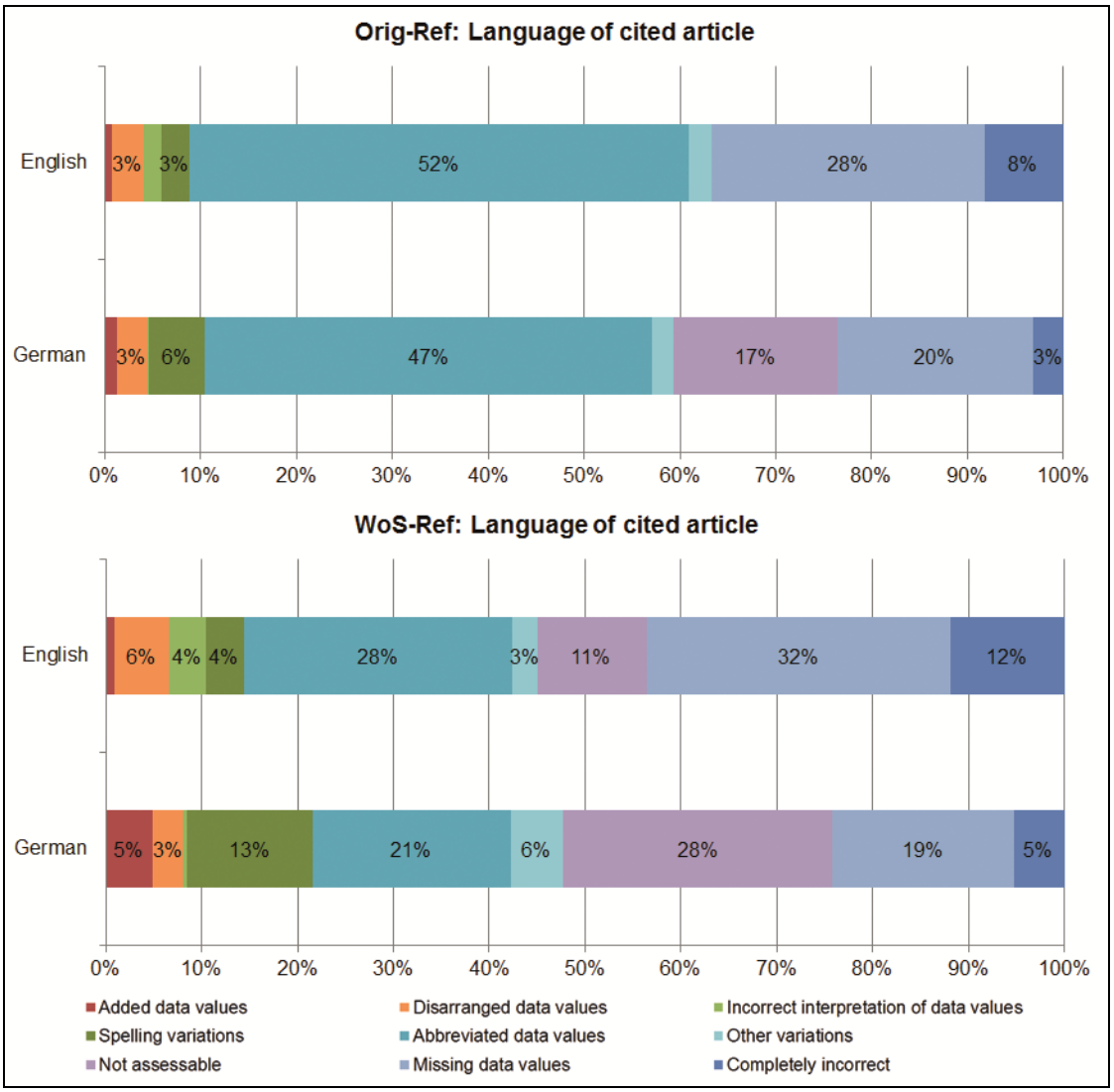


Figure 23: Inaccuracy subcategories per language of the cited article (source data values)

Other variations show a peak in references to German cited articles in the WoS-Ref result. This peak is primarily caused by German articles translated into English, which do not fully correspond to the translation in WoS and, therefore, were assessed as *Partially incorrect* (IAC *J*). The higher share of *Spelling variations* in the references to German cited articles in both result sets reflects the generally higher occurrence of *Special character*s (IAC *Q*) in languages other than English. The difference between the result sets again traces back to the above-mentioned handling of umlauts in WoS. The difference between the two languages in the category *Missing data values* is related to a general increase in *Omitted* ending pages and article titles in references to English cited articles and a particular increase in *Omitted* author-related data values, such as second initials. Furthermore, references to English target articles used *et al.* to shorten the list of author names in our data sample more frequently. The higher

share of *Completely incorrect* data values is due to a general increase in incorrect data values in references to English cited articles.

7.6 Evaluation per document type of the citing article

This section discusses the occurrences of inaccuracies according to the *document type of the citing article* (cf. Appendix F, Table 58-Table 65). Table 31 provides an overview of the distribution of citing articles over the document type categories. The columns on the left give the document types as classified in WoS; the columns on the right summarize the categories as merged in this evaluation. The normalized shares of inaccuracies per document type are quite equally distributed among the categories and lie between 11 and 22%, which is an indication that inaccuracies are not related to a specific document type. In the Orig-Ref result, the document type Letter is the most inaccurate (16%), while in the WoS-Ref result this is true for the document type Book Chapter / Book (22%). The document types with the lowest inaccuracy shares (both 11%) are Editorial Material in the Orig-Ref result and Other in the WoS-Ref result (cf. Appendix F, Figure 33).

Table 31: Overview of document type categories

Document types as classified in WoS		**Document types summarized for this evaluation**	
Article	3,039	Article	3,039
Review	464	Review	479
Review/Book Chapter	10		
Book Review	5		
Article/Proceedings Paper	191	Proceedings Paper	198
Proceedings Paper	7		
Editorial Material	98	Editorial Material	104
Editorial Material/Book Chapter	6		
Letter	54	Letter	54
Article/Book Chapter	36	Book Chapter / Book	51
Book	15		
Meeting Abstract	2	Other	4
Reprint	1		
News Item	1		

Figure 24: Inaccuracy subcategories per document type of the citing article (source data values)

Figure 24 illustrates the shares of inaccuracies per subcategory for all document types of the citing articles. The distribution of shares is fairly equal for the document types Article, Review, Proceedings Paper and Editorial Material in both result sets. Hence, the inaccuracies do not seem to be related to the document type. The other three document types draw a more differentiated picture, but the sample sizes are relatively small and do not allow us to draw valid conclusions. Therefore, we restrict the evaluation to a mere description of striking deviations from the other document types. In Letters, the share of *Disarranged data values* is the largest of all document types. At the same time they contain less *Missing data values*, but more *Completely incorrect* data values than other document types. Book Chapter / Book has

the highest share of *Spelling variations* and also the smallest share of *Abbreviated data values* in the WoS-Ref result. Additionally, it also has the largest share of *Not assessable* data values.

We conclude from these findings that the document type does not influence the inaccuracies made in references. While this result is more reliable for the document types with a higher number in our data sample, i.e. Article, Review, Proceedings Paper and Editorial Material, we refrain from making any further assumptions about the other document types. Supplementary research with a larger data sample may support or confute these results.

7.7 Evaluation per language of the citing article

This section summarizes the two result sets according to the *language of the citing article* (cf. Appendix F, Table 66-Table 72). 82% of all references come from English citing articles, 16% from German, 0.6% from French, 0.4% from Spanish citing articles and the remaining 1% comes from citing articles in other languages. The other languages are: Chinese, Croatian, Czech, Dutch, English/Spanish, Italian, Japanese, Korean, Lithuanian, Portuguese, Serbian, Spanish and Turkish[36]. The evaluation differentiates between languages with more than 10 citing articles and includes besides English and German also French and Spanish. The remainder are summarized as Other languages. However, the evaluation for languages other than English and German cannot be generalized, since the number of citing articles is too low. Analogously to the evaluation of document types, we restrict the evaluation to a description of notable deviations between the languages. Figure 25 details the shares of inaccuracies per language of the citing article summarized by the source data values.

The inaccuracies are quite evenly distributed over the different languages. In both datasets, citing articles in Other languages hold the most inaccuracies (around 25%). The next place with 22% of all inaccuracies is taken by French citing articles in the Orig-Ref sample, whereas in the WoS-Ref sample it is German citing articles with 23%. In contrast, German citing articles are the most accurate (15%) in the Orig-Ref sample. English and Spanish have a share of 18 and 17%, respectively, in the Orig-Ref result and both have 17% in the WoS-Ref result. French citing articles in the WoS-Ref result also have a 17% share. Hence, this finding corroborates the results of other studies that different linguistic backgrounds may lead to increased data inaccuracy in citing references (e.g. Moed, 2005). However, in our study this is only true for fringe languages, but not for German and Spanish.

[36] The number of citing articles per language is given in Table 67 in Appendix F.

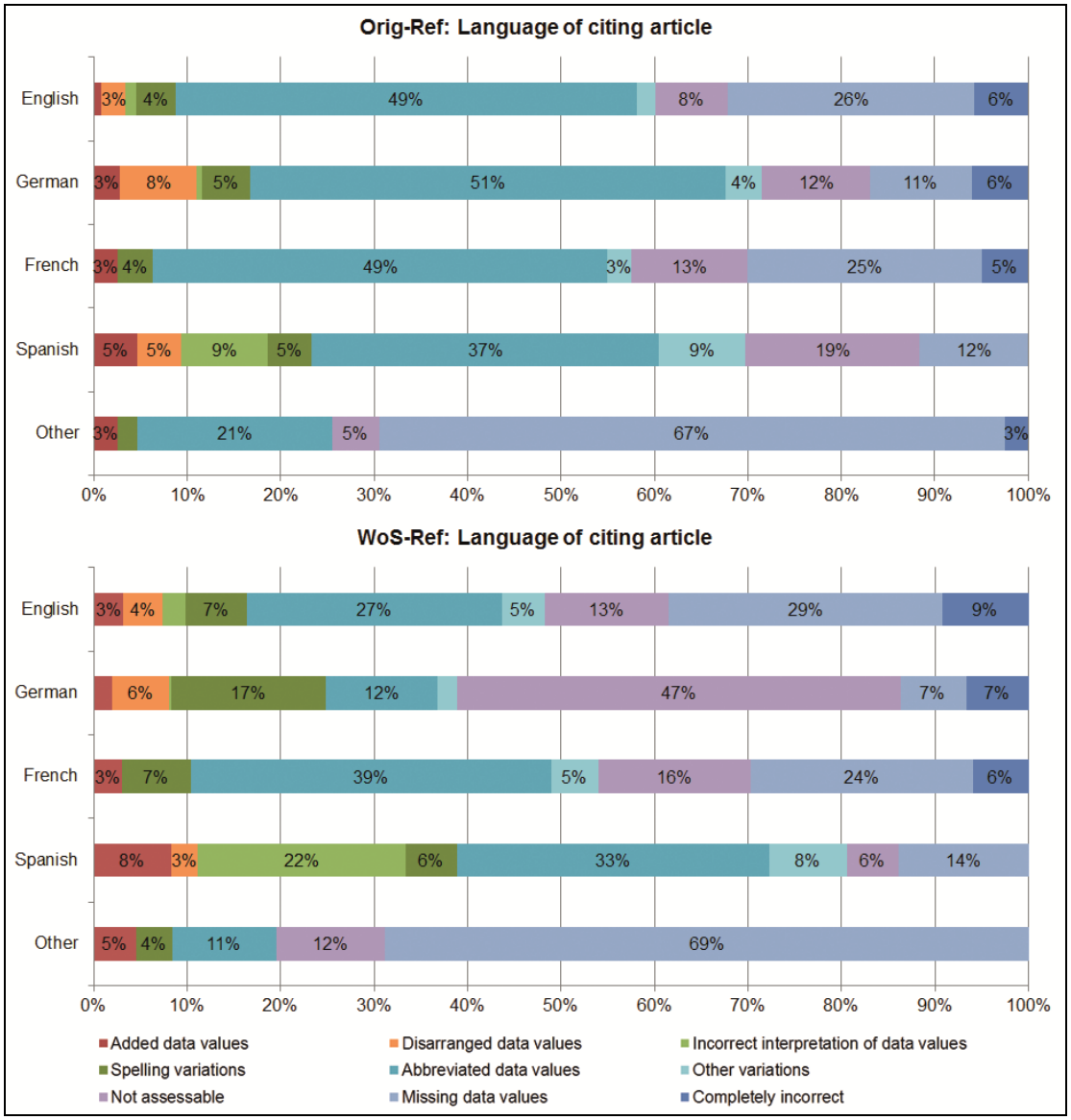


Figure 25: Inaccuracy subcategories per language of citing article (source data values)

In the context of languages, it is particularly interesting to analyze the language patterns of citing authors reflected in the language of the article title. While the majority of Spanish, French and Other language source articles cite English target articles, those that cite an article from the German dataset show a tendency to translate the article title into English (cf. *Not assessable*). Two-thirds of English source articles cite an English target article and 43% of those that cite a German one translate the German article title into English. In contrast, the majority of references from German source articles also cite German articles. However, 10% of them still cite them with a translated article title, which may suggest that citing authors have not checked or read the original article, but copied the reference from another bibliography.

Disarranged data values have a slightly higher share in German source articles, which originate from 11 references which cite the order of authors incorrectly (IAC *O*). *Spelling variations* are more common in references from German citing articles in the WoS-Ref result, since they pay more attention to the correct citation of umlauts. The relatively low share of *Abbreviated values* in the WoS-Ref result suggests that German citing authors cite the publication name and article title more accurately than English citing authors. Furthermore, *Missing data values* are less common in references from German citing articles, which indicates an overall increase in completeness of bibliographic data in references. In contrast, English and French citing articles have a larger share of *Missing data values*. In our data sample, Spanish citing authors have a higher share of the category *Incorrect interpretation of data values* in both data samples. Although compounded names are more frequent in the Spanish language, in our data sample they are cited by Spanish authors with greater variation. *Completely incorrect* data values do not occur at all. While references from citing articles in Other languages have the largest relative share of inaccuracies, they contain almost no *Completely incorrect*, no *Disarranged data values* and no *Other variations*. The majority of inaccuracies is caused by *Missing data values*.

7.8 Evaluation per citation window

This section documents the results of the quantitative analysis based on the different *citation window*s (cf. Appendix F, Table 73-Table 76). As explained in section 5.4, a variable citation window was chosen for all cited articles, resulting in references that were grouped into three five-year-citation windows: 1998-2002, 2003-2007 and 2008-2012. Consequently, the third citation window (2008-2012) has the highest absolute number of citing references, i.e. source records. However, since only the percentages based on the normalized inaccuracy shares are compared, these citation windows still indicate whether inaccuracies have changed over time. Both result sets show that the reference accuracy improves over time (cf. Figure 26). While 38% of all inaccuracies were detected in the first citation window, 33% were detected in the second and 29% in the third citation window.

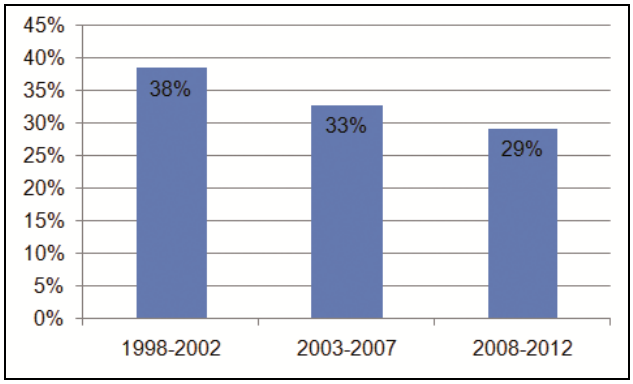


Figure 26: Shares of inaccuracies in the three citation windows for both assessment results (source data values)

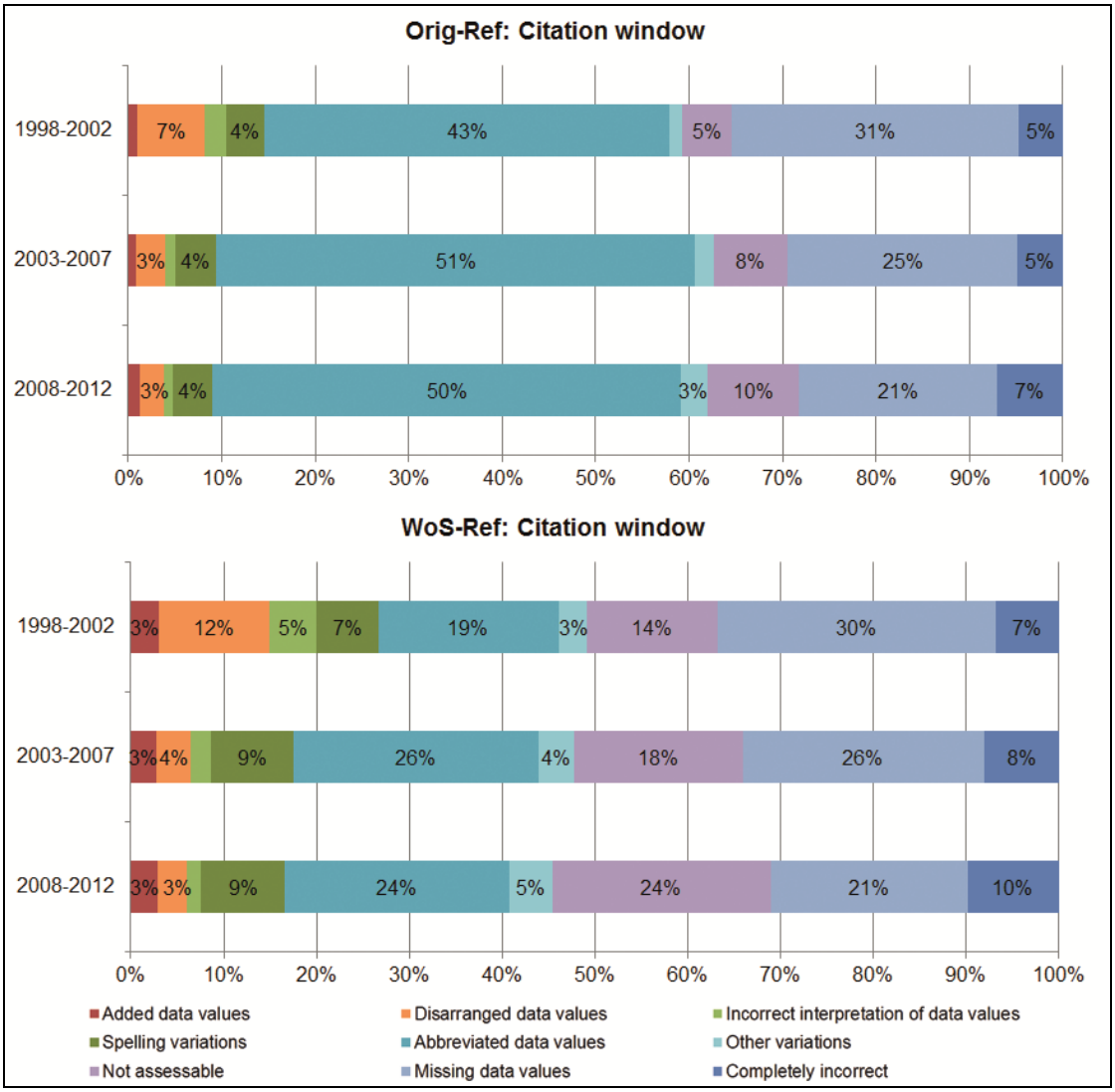


Figure 27: Inaccuracy subcategories per citation window (source data values)

The finding of increased accuracy over time continues in the evaluation of source data values and their decreased shares in the subcategories *Disarranged* and *Missing data values* for both assessment results (cf. Figure 27). The decrease in the *Disarranged data values* can be explained by the majority of references with a jumbled author order occurring in the first citation window (IAC *O*). Hence, the decrease can be related to a peculiarity in the data rather than to overall increased carefulness on part of the citing authors. The decrease in *Missing data values*, on the other hand, seems to stem from increased accuracy of citing authors. In the subcategories, *Added data values*, *Spelling variations* and *Incorrect interpretation of data values*, the shares of the different citation windows do not, or only slightly, change over time within each assessment result. Hence, we infer that these categories occur independently of time. In contrast, *Abbreviated data values*, *Not assessable*, *Other variations* and *Completely incorrect* increase over time. While we observed a decrease in the use of different *Abbreviation*s than the ISO abbreviation (IAC *I*) for publication names, an increase in *Cropped* (IAC *F*) values reflects an increased tendency to crop the ending page number in the reference. The increase in the *Not assessable* data values is, on the one hand, caused by missing ending page numbers in the WoS target records and missing volume numbers in the original articles. Since they occur in cited articles from 1998 and 2003, but not 2008, the citing references for these cited articles also increase over time. On the other hand, the increase reflects a general increased citation of German article titles both in German and in English translations (IAC *C Different language*). In the subcategory *Other variations*, all IACs increased slightly over time, thus, we could not pinpoint a single reason for the overall increased share. The increase in *Completely incorrect* data values must be attributed to the negligence of citing authors.

7.9 Evaluation of variants

As mentioned in section 5.3 and in the introduction of this chapter, variants of bibliographic fields in both assessment samples were assessed. Thus, the first step before the evaluation of the inaccuracy codes was to consolidate the variants into one result for each assessment sample. The assessment sample Orig-Ref contained one optional variant of the field *article title*. The assessment sample WoS-Ref contained two optional variants: one variant of the field *article title* (as described for the Orig-Ref sample) and one of the publication name. The publication names from the citing references were assessed against the full publication name as well as the ISO abbreviation of journal titles as recorded by WoS.

For both assessment samples, the variants were consolidated into one result set by choosing the most accurate version, i.e. for each record the result that contained fewer or minor inaccuracies was selected for inclusion in the final result set. In general, any other assessment results were chosen than the IACs *C Different language*, *I Abbreviation* and *Z Not available*, because these three IACs actually stand for an assessment that could not be continued. Other decisions that were taken in the process of consolidating the assessment results are documented in Table 32. The results of the consolidation of article titles are discussed in section 7.9.1; the consolidation of the publication names is discussed in section 7.9.2.

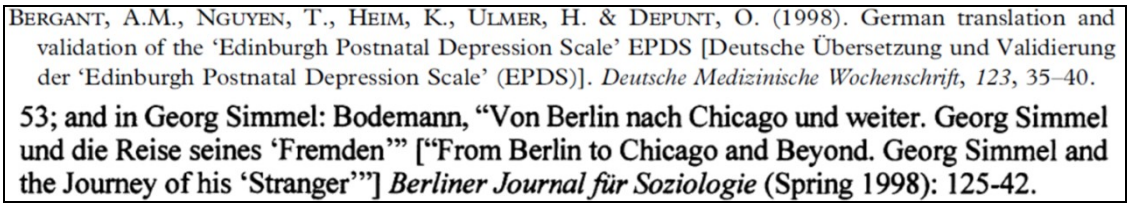
Table 32: Assessment decisions taken during the variant consolidation

Assessment result Variant A	Assessment result Variant B	Decision
B Q	B	Variant B - fewer inaccuracies
J	B	Variant B - involves less data manipulation
Q X	Q	Variant B - fewer inaccuracies

7.9.1 Evaluation of article title translations

31 references from the German dataset included an additional translation of the article title. In 25 of the references, the translated article title was in English, i.e. the article title cited in the reference was the German original with an English translation in brackets. In six cases, the main article title in the reference was the English translation, while the original German title was given as the translation[37]. As the goal in the assessment process was to obtain the least possible number of IACs *C Different language*, *I Abbreviation* and *Z Not available*, the German versions of the article title were used in the consolidated version in the Orig-Ref data sample (as the original article titles are in German), whereas in the WoS-Ref data sample the English versions were used (as WoS only provides English article titles). Figure 28 illustrates an example of both cases: the first reference shows an article title that has been translated into English and the original German article title is given in brackets; the second reference illustrates the reverse.

[37] Only one of these citing articles was German and gave the German original as the translation; all the others were English.


BERGANT, A.M., NGUYEN, T., HEIM, K., ULMER, H. & DEPUNT, O. (1998). German translation and validation of the 'Edinburgh Postnatal Depression Scale' EPDS [Deutsche Übersetzung und Validierung der 'Edinburgh Postnatal Depression Scale' (EPDS)]. *Deutsche Medizinische Wochenschrift*, *123*, 35–40.

53; and in Georg Simmel: Bodemann, "Von Berlin nach Chicago und weiter. Georg Simmel und die Reise seines 'Fremden'" ["From Berlin to Chicago and Beyond. Georg Simmel and the Journey of his 'Stranger'"] *Berliner Journal für Soziologie* (Spring 1998): 125-42.

Figure 28: Article title translations in two references

7.9.2 Evaluation of publication names and their abbreviations

In the WoS-Ref data sample the publication name was assessed against the full journal title, i.e. publication name, from the WoS target record (SO in WoS) as well as the ISO Source Abbreviation (JI[38] in WoS). In the consolidation process, it was found that 49% of all citations contained an abbreviated publication name and 40% contained the full publication name. 11% of the references held the same assessment result. Either this means that the publication name was the same or similar for both variants (cf. Table 33); or the journal title in the reference used an *Abbreviation* of the journal title which did not conform to the ISO abbreviation, thus, the assessment result of both variants was the IAC *I Abbreviation*. Of the references using an abbreviated journal title, 78% gave the correct ISO abbreviation of the publication name, 18% contained a different, but correct abbreviation of the journal title and 5% contained at least one inaccuracy.

Table 33: Similar publication name and abbreviations in WoS

Publication Name Original	Publication Name WoS (SO)	Publication Name Abbrev. WoS (JI)
Der Orthopäde	Orthopade	Orthopade
Hand Clinics	Hand Clinics	Hand Clin

7.10 False positive matches

False positive[39] matches are citing articles that show up in the citation count of a cited article, but do not, in fact, cite that specific article. The mismatch is usually caused by two references having the same or similar (depending on the citation matching algorithm) citation data that is used for the matching. Besides discovering 0.10% duplicates in the citing references of WoS,

[38] JI stands for ISO Source Abbreviation as opposed to J9 which stands for a 29-Character Source Abbreviation.

[39] The bibliographic data of all 33 false positive matches is given in Appendix G.

we also found 0.83% false positives. Two of these were corrections to an article, which we did not count as a citation, but categorized as false positive match. One cited article (from the German Sociology dataset) only had three false positive references and not a single correct one.

False positives occur equally in the NS and SSH. We found two particular peaks in the disciplines Sociology and Orthopedics (30% of all false positives each). The other disciplines each had one to five false positives. The majority of these citing articles was published in the first and second citation windows (1998-2007) and only 12% in the third citation window (2008-2012), which might be an indication that the citation matching algorithm in WoS had been changed over time. 90% of all false positives are English citing articles; the remainder are German. More than half of the false positives come from articles, 15% each come from Reviews and Proceedings papers. We also compared the domains in which the false positive citing articles were published with those of the cited articles. We found that 27% of false positives were not published in the same domain (NS/SSH) as their alleged cited article. Hence, it is unlikely that WoS incorporates the domain of the articles in the citation matching process.

7.11 Summary

This chapter presented the results of the quantitative analysis of inaccuracies identified in the qualitative content analysis. It answers the research questions about the frequency of inaccuracies and whether inaccuracy categories can be specifically related to one of the strata of the data sample. In general, the data accuracy of WoS target records is very high. Three-quarters of WoS records contain none to a maximum of two inaccuracies. This finding also corroborates the result of our pilot study (Olensky, 2013) in which we tested parts of the methodology employed in the present research. Most of the inaccuracies found in WoS records describe specific characteristics of the WoS data structure, such as the handling of special characters (e.g. German umlaut) and English article titles (independent of the original article language), but also trace back to inaccurate data extraction procedures.

Around 50% of all citing references contain one or two inaccuracies, while only a small share (around 15%) is discrepancy-free. However, on the data value level 85% of all assessed data values are completely accurate. The majority of inaccuracies in citing references occur in the category *complex* and *moderate* of the taxonomy described in section 6.3. Inaccuracies occur

most frequently as *Abbreviated data values*, *Missing data values* and *Completely incorrect* data values or data values that were *Not assessable* due to missing data in the target record. The majority of inaccuracies found in author-related fields are related to *Missing data values* and *Spelling variations*, reflecting *Omitted* initials, the use of *et al.* and the presence of a large number of German umlauts which can be transliterated differently. Article titles are predominantly *Omitted* due to the citation style, or cannot be assessed at all due to a language difference. With respect to publication names, the use of different *Abbreviation*s than the ISO abbreviations or of the full publication name causes the majority of inaccuracies. *Added data values* are the predominant inaccuracy subcategory in publication years, associated with the citation style, closely followed by *Other variations* which are caused by the IAC *T Plus/Minus*. The largest concern in volume numbers is *Disarranged data values* originating from issue numbers that were mistaken for the volume number. While starting page numbers are primarily *Completely incorrect*, ending page numbers tend to be *Cropped* (*Abbreviated data values*) or *Omitted*. In total, the most accurate bibliographic fields are publication year, volume number and starting page; the least accurate is the article title.

The distribution of the inaccuracy subcategories is quite similar in references to both NS and SSH target articles. However, we found that in the SSH fewer inaccuracies occur than in the NS, which in general reflects the difference in citation styles: NS tend to shorten the bibliographic data, whereas in the SSH all the bibliographic data is reproduced more accurately according to the original bibliographic data. In both assessment samples, the SSH disciplines have higher shares of accurate references, that are either non-discrepant or contain only one inaccuracy. Multidisciplinary Chemistry is the discipline featuring the most inaccuracies; Education & Educational Research the discipline with the fewest. *Disarranged data values* cannot be specifically attributed to a discipline or a respective citation style. While *Other variations* and *Completely incorrect* data values are more frequent in the SSH disciplines, *Missing data values* tend to occur more often in the NS disciplines and are caused by the prevailing citation practices.

The evaluation according to the *language of the cited article* revealed that references to German target articles contain more inaccuracies than references to English target articles. However, references to English cited articles contain larger shares of inaccuracies in the inaccuracy category *complex*, whereas the inaccuracies of references to German cited articles agglomerate in the categories *simple* and *moderate*. Hence, references to German articles may contain more inaccuracies, but they do not require as sophisticated data manipulation mechanisms to match them in a citation matching process as references to English articles.

In the different *document types of citing articles* we could not find any specific patterns of inaccuracies, i.e. the document type does not influence the types of inaccuracies made in references. They rather reflect the overall inaccuracy patterns.

The evaluation of the facet *language of the citing article* revealed a fairly equal distribution of inaccuracies among the different languages. However, references from citing articles in languages other than English, German, French and Spanish tend to be slightly more inaccurate. Interestingly, citing articles in languages other than English also have the highest shares of discrepancy-free records.

The evaluation of the three five-year-*citation window*s showed that reference accuracy increases over time, which indicates that authors have taken more care with their references in recent years. The decrease is mostly reflected in the two subcategories *Disarranged* and *Missing data values*, while *Completely incorrect* and *Not assessable* data values actually increase. Other inaccuracy subcategories do not vary much over time.

Even though the most accurate instances of the different facets have lower shares in the inaccuracy subcategories of the main group *simple*, such as *Added data values* and *Disarranged data values*, than in the subcategories of the main group *moderate* or *complex* (e.g. *Abbreviated data values*, *Missing data values* and *Completely incorrect*), we cannot infer from the mere occurrences of the inaccuracies that inaccuracies in the category *simple* are less likely to impact the citation matching process than inaccuracies in the category *complex*. Hence, in the following chapter we investigate the missed citations identified in the WoS *Cited Reference Search* in more detail and compare the matching capabilities of five other data sources (Scopus, GS, CWTS, iFQ and Science-Metrix) for these missed citations.

8 Evaluation of Missed Citations

In order to answer the questions what types of inaccuracies cause missed citations and how well the bibliometric data sources handle inaccurate data (RQ2; Part C and D of the applied methodology), the missed citations identified via the *Cited Reference Search* in WoS were analyzed and compared with the matching capabilities of Scopus, GS and the three applied bibliometric research groups CWTS, iFQ and Science-Metrix. In other words, we investigated which of the missed WoS citations could be matched by the other five data sources and which were missed by them as well. Section 8.1 sheds light on the strata in which missed citations in WoS occur. Section 8.2 compares the other five data sources with regard to the quantities of matched citations missed by WoS. Section 8.3 discusses the inaccuracies which caused the missed citations in detail and triangulates the data from all data sources to derive conclusions as to which inaccuracies impact the citation matching process. Section 8.4 summarizes the findings of the chapter.

8.1 Occurrences of missed citations in WoS

In total, 220 citations were identified in the *Cited Reference Search*, of which we could not obtain one citing article and the citation contained. Hence, the total number of missed citations investigated is 219. The overall missed citation rate (MCR) for our data sample in WoS accounts for 5.57%[40]. The MCR is lower in the NS (3%) than in the SSH (10%), which is also reflected in the MCRs of the disciplines (cf. Table 34). Political science shows the highest share of missed citations (12%), while the shares of the NS disciplines are almost equally low (3-4%). The MCRs of references to English (6%) and German (5%) cited articles are almost the same and they are also almost constant over time (1998-2002: 6%; 2003-2007: 5%; 2008-2012: 6%). 77% of missed citations were found in English citing articles, followed by 21% in German citing articles. The remaining missed citations are attributed to 3 French citing articles and 1 Spanish citing article. Yet again, the MCR of each citing article language does not vary

[40] This percentage does not include the one missed citation we could not verify.

greatly: it is almost equally high for English, German and Spanish citing articles (5-7%) and slightly higher for French citing articles (13%). 82% of missed citations were retrieved from Articles, 8% from Proceedings Papers, 5% from Reviews and 4% from Book Chapter / Book. Editorial material and Letter have the smallest shares of missed citations. The MCRs vary more according to the different document types than to other facets of the data sample, with Book Chapter / Book and Proceedings Papers having the highest and Reviews and Letters the lowest rates (cf. Table 34).

Table 34: Missed citation rates per discipline and document type

Discipline	MCR within each discipline	Document type	MCR within each document type
Multidisciplinary Chemistry	3%	Article	6%
Internal & General Medicine	3%	Review	2%
Orthopedics	4%	Proceedings Paper	9%
Education & Educational Research	8%	Editorial Material	4%
Political Science	12%	Letter	2%
Sociology	8%	Book Chapter / Book	16%

Thus, we can conclude that the domain, discipline of the cited article and document type of the citing article influence the occurrences of missed citations. Since the citation style is closely related to all three characteristics, this suggests that the citation style impacts the match and non-match of citations.

8.2 Comparison of missed citation matches by Scopus, Google Scholar, CWTS, iFQ and Science-Metrix

This section compares the efficiency of the different matching algorithms of Scopus, GS, CWTS, iFQ and Science-Metrix in handling citations missed in WoS. As mentioned section 2.4.3, all three applied bibliometric research groups work with raw data from WoS which they match according to their citation matching algorithms developed in-house. Science-Metrix additionally uses Scopus data for its analyses, because the Scopus raw citation data provides article titles which Science-Metrix incorporates into its citation matching process. However, to

ensure comparability with the other two institutions only the data that was matched, based on the cited reference information from WoS, was used. Citing articles which were not covered in a data source were excluded from the calculation. In the course of the evaluation, we additionally found that Scopus did not cover six cited articles and GS did not cover one (even though the journals were available in the databases). We decided not to exclude these cited articles in general, but to add the corresponding citing articles to the count of non-covered citing articles. The numbers of citing (and cited) articles not covered in each data source are listed in Table 35.

Table 35 gives an overview of how many of the missed WoS citations were matched in the other data sources. The first row, *not covered*, gives the number of missed WoS citations where the citing articles were not covered in the data sources and, therefore, could not have been matched correctly. Based on this, the second row, *potential matches*, lists the number of citations that could potentially have been matched (219 missed WoS citations minus the non-covered citing articles). The next row, *matched*, lists the number of citations which the respective data source was able to match correctly (of the potentially matchable number of citations in row two). The fourth row, *missed*, gives the number of citations that could not be matched, but were covered by the data source. In other words, these missed citations potentially contain inaccuracies which led to the non-match in the citation matching process. The last row, *matched %*, gives the respective percentages of missed WoS citations which were correctly matched by the data source (the base is the row *potential matches*).

95% of the missed WoS citations are also not matched in the Science-Metrix data, i.e. Science-Metrix could only match 5% of the missed WoS citations. Scopus and GS, on the other hand, were able to match more than half of the missed WoS citations and Scopus performed slightly better than GS. iFQ matched more missed WoS citations (79%) than CWTS (76%). The *plus 5* matched citations in the iFQ column stand for five references which were denoted as uncertain matches and would require manual effort to verify the citations. In all five cases the matching was correct and, therefore, added to the total number of citations matched by iFQ. However, without these five uncertain matches, the iFQ's result would be only one matched citation ahead of CWTS. In this comparison, it is also important not to overlook the shares of not covered references (in relation to the total number of missed citations in WoS), which are the highest for Scopus (12%), followed by CWTS, iFQ and Science-Metrix (all 4%) and GS (2%). As CWTS, iFQ and Science-Metrix work with WoS data, the non-covered documents are the same (all books or book chapters), except for one additional article which was not covered in the Science-Metrix database.

Table 35: Comparison of the data sources – missed citations

	WoS	Scopus	Google Scholar	CWTS	iFQ	Science Metrix
not covered	-	27	5	8	8	9
potential matches	-	192	214	211	211	210
matched	-	134	135	160	161 (plus 5)	11
missed	219	58	79	51	45	199
matched %	-	67%	63%	76%	76% (79%)	5%

Four references (cf. Appendix I, Table 84) were missed by all matching algorithms, although covered by all data sources. None of them contain a volume number (IAC *E Omitted*). One reference additionally features a jumbled order of author names for the first three authors (IAC *O Incorrect order of authors*). The other three references not only lack the volume number, but also the starting page. One of them additionally misses a prefix in the last name of the first author (IAC *J Partially incorrect*). The other two do not contain a publication year, use the correct ISO-abbreviated publication and have the additional information (IAC *N Additional information*) *in press* or the German *im Druck* in the publication name. Hence, they are references to a forthcoming publication, which explains the missing bibliographic information (publication year, volume number, pagination) which would have been needed to correctly match these citations.

Limitations. As it would go beyond the scope and the resources of this dissertation, we only compared the citations that were missed by WoS, but did not perform an overall comparison of overlap and coverage of all citations (matched, missed and false positive). Therefore, we cannot calculate a valid overall missed citation rate for each data source. In future work, we will investigate and compare the data sources according to matched and missed citations as well as false positives (cf. section 10.2). Furthermore, Science-Metrix usually uses Scopus and WoS citation data for its citation matching, therefore, its performance in the WoS citation matching process may not reflect the true ability of its algorithm. Moreover, GS's relatively good performance should not lead us to disregard its many (often reported) data quality problems, such as false positives, duplicates, etc. (e.g. Harzing, 2008; Jacsó, 2005a, 2005b, 2005c). A higher recall of missed WoS citations does not necessarily mean that the precision of other matched citations is higher as well.

8.3 Analysis of inaccuracies in missed citations

This section focuses on the inaccuracies identified in the missed citations. Section 8.3.1 discusses the inaccuracies in missed citations by WoS according to the three assessment results Orig-Ref, WoS-Ref and CitedRef-WoS. Section 8.3.2 describes the data triangulation with the five other data sources, based on the following assessment results: the inaccuracies of cited reference information in Scopus (CitedRef-Sco) of the citations missed by WoS and Scopus, the inaccuracies of missed citations GS could not match, based on the Orig-Ref and WoS-Ref data, and the inaccuracies of missed citations the three applied bibliometric research groups were not able to match, based on the CitedRef-WoS data. By identifying the inaccuracies present in the missed citations and pinpointing the IACs which were solely responsible for the non-match it will be possible to determine the IACs with the greatest impact on the citation matching process (cf. section 8.3.3). Appendix I gives the results of all assessment sets (cf. Table 84-Table 92).

The GS data was based on the Orig-Ref and WoS-Ref assessment results. All inaccuracies were inherited from the Orig-Ref result, except for the IAC *Z Not available* in the volume numbers and the IAC *I Abbreviation* in the publication names. Since they mark data values that could not be further assessed in the Orig-Ref assessment process, every occurring IAC *Z Not available* and IAC *I Abbreviation* in publication names in the Orig-Ref result was replaced by the respective assessment result of the WoS-Ref dataset. Five of the citations missed in GS did not contain a discrepancy in the original references. Hence, we can infer that these were missed due to inaccuracies caused by the data extraction and handling. Scopus missed a total of 58 citations which were not matched by WoS. Of these, six citing articles did not contain cited reference information at all and four citing articles did not contain the cited reference information to the cited article in question. Hence, 10 citations were not matched solely on account of missing cited reference information and could not be assessed for the occurrence of inaccuracies.

We excluded the inaccuracy subcategory *Not assessable* from the data analysis of the missed citations. The IAC *Z Not available* refers to data values which were not available in the verification data source and does not reflect inaccuracies in the references themselves. The IAC *C Different language* describes the specifics of the data structure in WoS (cf. chapter 7),

thus, it only occurs in article titles, which were not part of the citation matching processes employed by the majority of bibliometric data sources compared[41].

8.3.1 Analysis of WoS missed citations

In order to identify the inaccuracies which caused the non-match in WoS, we assessed the cited reference information of the missed citations against citations which were matched correctly, i.e. CitedRef-WoS assessment (cf. section 5.2, Part C). The occurrences of inaccuracies were analyzed according to the subcategories of the taxonomy of bibliographic inaccuracies and compared to the Orig-Ref and WoS-Ref result sets. In contrast to chapter 7, only the bibliographic fields occurring in the cited reference information (author last name, first and second initial, publication name, publication year, volume number and starting page) were examined, because the reason for the non-match must consequently lie in one (or more) of these bibliographic fields. The counts of inaccuracies were again normalized by the number of IACs present in the respective subcategory, but not by the number of values assessed, since these were the same for all three datasets (Orig-Ref, WoS-Ref and CitedRef-WoS). The inaccuracies identified in the CitedRef-WoS assessment process were not the same for every record as those identified in the Orig-Ref and WoS-Ref assessment process. Most of the inaccuracies refer to mistakes typically made by researchers when they compile their bibliographies. However, some of the inaccuracies were assessed as such because WoS had introduced additional inaccuracies. Hence, we compared the occurrences in the three assessment results with each other and consequently pinpointed which inaccuracies were caused by

- authors and were not corrected by WoS – could be responsible for the non-match
- authors, but were corrected by WoS – not responsible for the non-match
- the citation style, and not handled appropriately by WoS – could be responsible for the non-match
- the data extraction or handling process by WoS, or inaccurate data provided by journal publishers which was not corrected by WoS – could be responsible for the non-match.

[41] Since 88% of the missed citations in Scopus, contained a correct article title and were still not matched correctly, we concluded that Scopus does not use it in its citation matching. We cannot make a clear statement about the matching in GS, but a comparison of the matching results suggests that GS employs the article title in its citation matching process.

In the CitedRef-WoS result, the shares of inaccuracies in the bibliographic data fields assessed are as follows: 28% of all inaccuracies were identified in the starting page, 20% in the volume number, 13% in the first author's last name and 13% in the first initial, 10% in the publication name, 9% in the publication year and 7% in the first author's second initial. The shares of inaccuracy subcategories are displayed in Figure 29 (third bar from the top) and discussed in the following paragraph.

The most inaccuracies in the CitedRef-WoS result are *Completely incorrect* starting pages. A closer examination showed that 51% of them actually held the cited page number. Another six CitRefmiss source records also held the cited page number, but could have been converted into the correct starting page (assessed as IAC *T Plus/Minus* or IAC *H Jumbled value*). In total, the inaccuracy subcategory *Completely incorrect* is the most frequent of all subcategories. *Completely incorrect* values occur in volume numbers and also, in small quantities, in all other four bibliographic fields, except for the author's last name. The second most frequent inaccuracy subcategory is *Missing data values*. They occur almost equally in the volume number and the starting page (IAC *E Omitted*), followed by the author's second initial (IAC *E Omitted* and IAC *P No author name*), and a few times in the publication year as well as in the author's first initial and last name. Next in the ranking come *Other variations*, which are most common in the publication year, followed by the starting page and volume number. In these typically numerical fields, although the volume number can sometimes be a string, the correct value could be calculated by manipulating one digit or the entire number (IAC *T Plus/Minus*). The fourth most frequent inaccuracy subcategory is *Disarranged data values*, in which the inaccuracies are mainly caused by issue numbers substituted for volume numbers or starting pages (IAC *G Interchanged fields*). A *Jumbled value* (IAC *H*) or an *Incorrect order of authors* (IAC *O*) was the reason for the remaining inaccuracies identified in this subcategory. Three subcategories rank fifth: *Added data values*, *Incorrect interpretation of information* and *Abbreviated data values*. Inaccuracies in the subcategory *Added data values* were mainly caused by additional *Punctuation* in the author's first name and last name (IAC *R*), *Additional information* in the publication name (IAC *N*), and references which gave the author's first names in full (IAC *U Full first name*). *Incorrect interpretation of information* was only caused by the IAC *M Incorrect interpretation of author names*, which occurred in six CitRefmiss source records. *Abbreviated data values* were mainly found in the publication name (IAC *I Abbreviation*). The latter contained either a different *Abbreviation* than the CitRefmatch source record or used the full publication name or seemed to be cropped after the field's character limit. A few *Cropped* volume numbers and starting pages (IAC *F Cropped*) also contributed to the total number of inaccuracies in this subcategory. *Spelling variations* rank last, occurring

only in the field author's last name. The prevalent IAC in this subcategory is the IAC *B Spelling error*. The IAC *Q Special character* decodes a few inaccuracies, where a Germanic umlaut has not been processed correctly by WoS.

On comparing the inaccuracies of the CitedRef-WoS with the Orig-Ref and WoS-Ref results, we discovered that around 30% of all inaccuracies found in the original references (of the 219 missed citations) were corrected in the data handling and citation matching process of WoS. In the case of the Orig-Ref result, this mainly concerned *Abbreviated data values* stemming from abbreviated publication names (IAC *I*). However, 10% of divergent publication names remained in the CitedRef-WoS result and, therefore, are still likely to have had an impact on the non-match of the specific citation. Moreover, more than 30% of *Omitted* second initials were corrected by WoS compared to the Orig-Ref result and 14% of *Omitted* second initials compared to the WoS-Ref. Hence, we infer that a matching second initial is not crucial in the WoS citation matching process. In comparison to the WoS-Ref result, around 50% of *Spelling variations* in authors' last names, consisting of *Spelling error*s (IAC *B*) and *Special character*s (IAC *Q*), were corrected as well as all *Spelling variations* in the publication name. Therefore, we conclude that *Spelling variations* in the publication name can be handled by the WoS citation matching process well. We can only assume that this is due to the fact that the matching process employs a list of variations, i.e. ISO abbreviations and WoS-specific abbreviations of the journal names, allowing for more flexibility in the matching.

However, WoS not only corrected inaccuracies, but also introduced additional inaccuracies or changed the type of assessment result. In total, 57% of inaccuracies still present in the CitedRef-WoS result were caused by authors, 12% were due to the citation style which WoS was not able to process correctly, and 31% of inaccuracies were traced back to inaccuracies introduced in the data handling process or to data which was already inaccurately provided to WoS[42]. Inaccuracies caused by authors were the most frequent in the starting page, volume number and publication year. Fewer inaccuracies occurred in authors' last name and second initial. They are primarily attributed to *Missing data values*, *Other variations* and *Disarranged data values*. The non-matches caused by the citation style almost all refer to cited page numbers which were perceived as starting pages. In total 14% of all CitRefmiss source records contained the cited page number[43], which were predominantly assessed as *Completely*

[42] 16 references did not contain a single inaccuracy, but were still not correctly matched by WoS. They are given in Appendix I, Table 95.

[43] However, not all of the cited page numbers in the cited reference information in WoS were caused by the citation style, as some of the original references also contained the correct pagination and only gave the cited page number in addition. These inaccuracies were counted as caused by WoS.

incorrect values. The inaccuracies introduced by the data handling process chiefly occurred in the authors' first initials, last names, publication name and starting page, which are quite evenly attributed to *Added data values*, *Missing data values* and *Incorrect interpretation of data values*.

In 47% of the missed citations, a single inaccuracy was responsible for the non-match. Table 93 in Appendix I gives an overview of the inaccuracies which definitely caused a non-match in the CitedRef-WoS result, because no other inaccuracy co-occurred. Of those citations, 49% were missed because of a discrepancy in the starting page, 30% due to a discrepant volume number, 15% because of a discrepancy related to an author name, whereof last names were the most inaccurate, and 6% because of an inaccurate publication year. The prevailing inaccuracy subcategory is *Completely incorrect*, followed by *Disarranged data values* and *Other variations*. Even though we found that WoS was able to handle *Spelling variations* in the publication name well, this was not the case for *Spelling variations* in the author's last name. Although this result shows that a discrepant second initial or publication name was not solely responsible for a non-match and both bibliographic fields tend to be corrected by the data handling process in WoS, we cannot simply infer that they do not influence the citation matching process at all. In our data sample they co-occur with other inaccuracies, therefore, we can only assume that the second initial plays an inferior role in the citation matching process to the publication name.

8.3.2 Comparison with Scopus, Google Scholar, CWTS, iFQ and Science-Metrix

We compared these findings to the occurrences of inaccuracies in citations missed by the other five data sources. The goal was, on the one hand, to verify the results of the analysis of missed WoS citations, and, on the other hand, to determine, where possible, differences in the ability of the matching algorithms to handle inaccurate data. Figure 29 gives an overview of the inaccuracy subcategories identified in all six data sources. The first two bars refer to the assessment results of the original references against the original target article (Orig-Ref) and against the WoS target records (WoS-Ref), but are restricted to the missed citations and the bibliographic fields used in the WoS citation matching process. The third bar shows the results of the CitedRef-WoS result. The remaining five bars display the results of the other five bibliometric data sources: GS, with results based on the Orig-Ref and WoS-Ref result; Sco, with results based on the inaccuracies found in the cited reference information in Scopus (CitedRef-Sco); and the three applied bibliometric research groups, each result based on the

data of the CitedRef-WoS result. The inaccuracy subcategories for all five data source in comparison to the CitedRef-WoS result are discussed in the order of display in Figure 29.

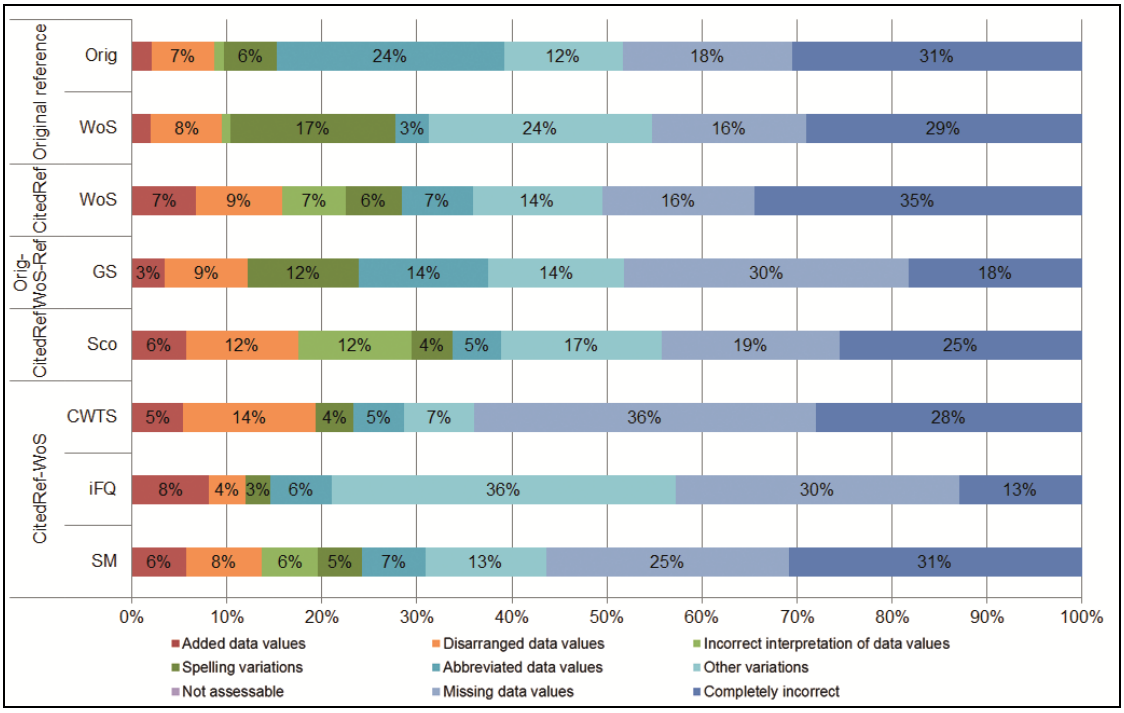


Figure 29: Comparison of inaccuracy subcategories in missed citations for each data source

Added data values have the largest share in the iFQ data, closely followed by Scopus and Science-Metrix. In the GS, Scopus, CWTS and iFQ data, they are predominantly caused by the IAC *N Additional information*, which refers to the addition *in press* of forthcoming articles. A few instances of IAC *N* also denote *Additional information* about the issue (e.g. AUG, FEB). However, in spite of *Added data values*, the data sources also matched some references correctly, as long as either one or a combination of the fields publication year, volume number and starting page were not missing. In the Science-Metrix data, almost the same number of *Added data values* occurs as in the CitedRef-WoS result, only the figures for the IAC *R Punctuation* and *U Full first name* are lower.

Disarranged data values have the largest share in the CWTS data. The majority occurs in the volume number, which holds the issue number or the starting page, followed by the starting page number holding the issue number (IAC *G Interchanged fields*). With the other data sources the picture is the same, except for iFQ, where only two issue numbers in the place of a volume number were not corrected. The remainder of the numerical *Disarranged data values* was correctly matched. Two records, in which the author order of the first and second authors

(IAC *O Incorrect order of authors*) was jumbled, were missed by Scopus, CWTS, iFQ and Science-Metrix, while one of these records was matched correctly by GS.

The subcategory *Incorrect interpretation of data values* only arises in Scopus and Science-Metrix data. In both data sources, the responsible inaccuracy is an *Incorrect interpretation of author names* (IAC *M*). While Scopus introduced additional inaccuracies during its data extraction and handling process, Science-Metrix did not match the inaccuracies caused by WoS. The matching algorithms of CWTS and iFQ were able to match these records correctly. In the original reference, only one record contained the IAC *M* and this was correctly matched by GS.

Spelling variations have the largest share in GS. Citations missed by GS contain both *Spelling error*s (IAC *B*) and *Special character*s (IAC *Q*), whereas citations missed by Scopus, CWTS and iFQ do not contain any *Special character* discrepancies and only a few uncorrected *Spelling error*s. Science-Metrix was able to correct a few *Spelling error*s, the remainder of the *Spelling variations* are the same as in the CitedRef-WoS result.

GS also has the largest share of *Abbreviated data values*, whereas the other data sources have a share between 5 and 7%. The subcategory consists of abbreviated publication names (IAC *I*) and very few *Cropped* starting pages and volume numbers (IAC *F*) in all data sources. The exception was Scopus, whose matching process we found to be very robust to variations of publication names, comparing examples of correctly matched publication names to those present in missed citations.

The iFQ data shows the largest share of *Other variations*, which are mainly caused by publication years and only by a few starting pages which could have been calculated correctly (IAC *T Plus/Minus*). The shares are lower for Scopus, GS and Science-Metrix, while CWTS has the lowest share. In the other data sources, *Other variations* also occur in the publication year and starting page and, in the Science-Metrix data, they also occur in the volume number.

In two of the five data sources, namely GS and CWTS, the prevailing inaccuracy subcategory in missed citations is *Missing data values*. For the other three data sources, Scopus, iFQ and Science-Metrix, *Missing data values* rank second. While in GS the inaccuracies are mainly caused by *Omitted* starting pages, volume numbers and a few missing author names due to the citation style (IAC *P No author name*), all but one of the IACs *P* have been matched in the CitedRef-WoS and the CitedRef-Sco data. Therefore, in all other data sources the inaccuracies in this subcategory consist of *Omitted* starting pages, volume numbers and publication years.

Scopus and Science-Metrix have the highest share of the subcategory *Completely incorrect*, GS and CWTS the next highest, while iFQ has the lowest share. In GS, all but one of the *Completely incorrect* values occur in the starting page. The other instance is found in the publication year. In the other data sources, the majority of *Completely incorrect* data values occurs in the starting page, followed by the volume number. Fewer occurrences are found in the publication year, and only in the Scopus and Science-Metrix data in the second initial. Hence, all incorrect volume numbers were handled by GS's matching algorithm.

Summarized differently, Figure 30 gives an overview of which IACs occur in missed citations in all six data sources. Since the basis of the comparison were the 219 missed citations and the inaccuracies identified in the CitedRef-WoS result, this dataset contains the highest number of inaccuracies for each IAC in all but two cases (except for IAC *Q Special character* and *P No author name* in GS which is explained in the following paragraph). Therefore, comparing the absolute numbers of each IAC with the other data sources enables us to determine what kinds of inaccuracies the citation matching algorithms of the other data sources are able to handle compared to WoS. The IACs on the x-axis are sorted by the frequency in the CitedRef-WoS result. The ranking of the data sources follows the ranking of IACs in the CitedRef-WoS result.

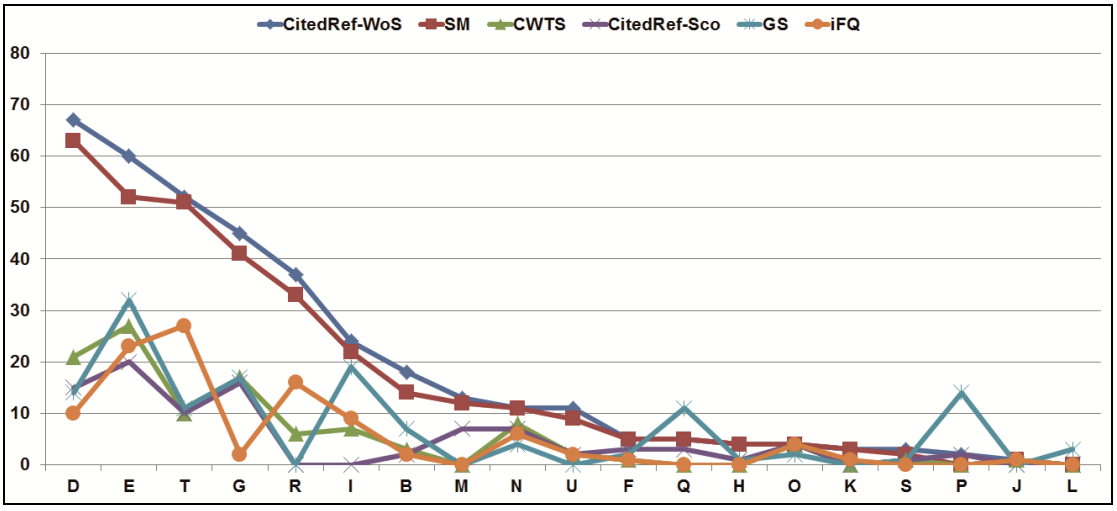


Figure 30: IACs occurring in the data values of missed citations (absolute numbers)

The biggest problems in missed citations in WoS are the IACs *D Completely incorrect*, *E Omitted* and *T Plus/Minus*. The number of inaccuracies in the Science-Metrix missed citations is distributed analogously to the CitedRef-WoS inaccuracies. However, Science-Metrix handled a few, but not all of the IACs *D Completely incorrect*, *E Omitted*, *G Interchanged*

fields, *R Punctuation*, *I Abbreviation*, *B Spelling error*, *M Incorrect interpretation of author names* and *U Full first name*. Furthermore, Science-Metrix correctly matched all citations with IAC *P No author name*. Hence, the problem areas are the same as in WoS. CWTS handled the least number of *Completely incorrect* data values (IAC *D*) and *Additional information* (IAC *N*) from WoS compared to the remaining data sources, Scopus, GS and iFQ. However, the biggest problem area for CWTS is *Omitted* data values (IAC *E*). CWTS matched all citations containing the IACs *M Incorrect interpretation of author names*, *Q Special character*, *H Jumbled value*, *K Space* and *P No author name*. In the Scopus data, the biggest issues are the IACs *E Omitted*, *G Interchanged fields* and *D Completely incorrect*. However, the CitedRef-Sco result contained the least *Omitted* data values of all data sources, suggesting that its matching algorithm might be more robust against this kind of inaccuracy. Scopus is the only data source, apart from Science-Metrix, that was not able to match all citations containing the IAC *M Incorrect interpretation of author names*. Analogously to WoS, Scopus was also not able to handle all citations with IAC *P No author name*. GS has the same number of *Completely incorrect* data values (IAC *D*) as Scopus, which is less than CWTS, but more than iFQ. The highest number of inaccuracies is attributed to *Omitted* data values (IAC *E*) in GS. Since the GS result is based on the Orig-Ref and the WoS-Ref result, it also reflects the non-assessed aspects in the processes, *Punctuation* (IAC *R*) and *Full first name* (IAC *U*), as well as the comparably high figures for *Abbreviated* values (IAC *I*), *Special character*s (IAC *Q*), *No author name* (IAC *P*) and *Informational letter*s (IAC *L*). None of the data sources, except GS, were able to correct the *Incorrect order of authors* (IAC *O*) in two out of four missed WoS citations. For iFQ the biggest concern is the IAC *T Plus/Minus*. The algorithms of Scopus, GS and CWTS work better in counterbalancing values in which one or two digits have been transposed. However, iFQ handled *Completely incorrect* data values (IAC *D*) best and *Omitted* data values (IAC *E*) second best. While Scopus, GS and CWTS have around the same number of *Interchanged fields* (IAC *G*), iFQ's algorithm was able to deal with almost all of them. Analogously to GS and CWTS, iFQ was also able to match all citations with an IAC *M Incorrect interpretation of author names*. Furthermore, citations with *No author name* (IAC *P*) were matched as well. Like CWTS, iFQ handled all *Jumbled value*s (IAC *H*) and *Special character*s (IAC *Q*) and was the only source which could also deal that all citations containing *Padded* values (IAC *S*).

8.3.3 Data triangulation with Scopus, Google Scholar, CWTS, iFQ and Science-Metrix

It should be borne in mind that, just because an inaccuracy occurs in the cited reference information, it does not imply it was also responsible for the non-match. When inaccuracies co-occur it is difficult to pinpoint a sole reason for the non-match without knowing the specific structure of the matching algorithm. For example, a citing reference may hold a *Completely incorrect* first initial for the first author (IAC *D*) and at the same time the publication name may hold *Additional information* (IAC *N*), such as AUG for the issue number. Since we do not know the weight of a match of the two fields in the matching algorithm, it is not possible to determine whether the non-match was caused by a single inaccuracy, which led to the abortion of the matching of the specific citation, or whether the non-match was caused by a combination of inaccuracies. To determine individual IACs which were responsible for the non-match in each data source, we triangulated the assessment results of missed citations containing only one inaccuracy. As discussed in section 8.3.1, in 47% of the WoS missed citations just one inaccuracy was responsible for the non-match. This was the case for 43% of all citations missed by GS, 36% by Scopus, 31% by CWTS, 18% by iFQ and 21% by Science-Metrix[44]. Table 36 summarizes the IACs responsible for the non-matches in the data sources.

Table 36: Single occurrence of the inaccuracies in a reference caused a non-match

	CitedRef-WoS	GS	Scopus	CWTS	iFQ	SM
B Spelling error	x	-	-	-	-	x
D Completely incorrect	x	x	x	x	x	x
E Omitted	x	x	x	-	-	x
F Cropped	x	x	x	-	-	x
G Interchanged fields	x	x	x	x	-	x
H Jumbled value	x	-	x	-	-	-
I Abbreviation	-	x	-	-	-	-
M Incorr. interpret. ANs	x	-	-	-	-	-
Q Special character	x	x	x	-	-	x
R Punctuation	x	-	-	-	-	x
S Padded	x	-	-	-	-	x
T Plus/Minus	x	x	x	x	x	x
U First full name	x	-	-	-	-	-

Completely incorrect (*D*) and *Plus/Minus* (*T*) are the two IACs which definitely led to a missed citation in all data sources. *Interchanged fields* (IAC *G*) caused a missed citation in

[44] Since the data sample was reduced for this evaluation, the results represent tendencies, not final evidence. They need to be corroborated by further studies discussed in section 10.1.

five out of six data sources. Hence, these three IACs are the biggest threats to a successful citation matching. *Omitted* (IAC *E*) and *Cropped* (IAC *F*) data values led to a non-match in four out of six data sources, i.e. a single *Omitted* or *Cropped* data value did not hinder a correct match in the data of CWTS and iFQ. However, since the assessment results of the two data sources contain the IACs *E* and *F*, it can be inferred that, if the *Omitted* or *Cropped* value co-occurs with another inaccuracy, the citation may fail to be matched. *Special character*s (IAC *Q*) also caused a missed citation in four out of six data sources. In this case, however, CWTS and iFQ were able to handle all *Special character*s and they do not occur in any missed citation. All other IACs are specific to the data source. While *Spelling error*s (IAC *B*), *Punctuation* (IAC *R*) and *Padded* values (IAC *S*) can cause missed citations in WoS and Science-Metrix, *Jumbled value*s (IAC *H*) do so in WoS and Scopus. *Abbreviation*s (IAC *I*) can be responsible for a non-match in GS, whereas an *Incorrect interpretation of author names* (IAC *M*) or a *Full first name* (IAC *U*) can only cause a citation to be missed in WoS.

To summarize the findings of the data triangulation, we categorized the IACs according to their impact on the citation matching process:

- IACs not occurring at all in a data source have *no impact*, since a non-occurrence cannot cause a missed citation.
- IACs which were identified as the sole reasons for the non-match of the citation were categorized as having an *explicit impact*.
- The remaining IACs were categorized as having a *potential impact*.

The matrix in Figure 31 lists all IACs ordered by the inaccuracy subcategories for all three categories *no impact*, *potential impact* and *explicit impact*. The *x* marks each IAC in one of the three categories for each data source. That way it describes whether a specific IAC had *no impact*, *potential impact* or *explicit impact* on the citation matching process of a data source.

Space	K							Explicit impact
Informational letter	L							
Additional information	N							
Punctuation	R	x					x	
Padded	S	x					x	
Full first name	U	x					x	
Interchanged fields	G	x	x	x	x		x	
Jumbled value	H	x	x					
Incorrect order of authors	O							
Incorrect interpretation of names	M	x						
Incorrect interpretation of additional information	V							
Typographical variation	A							
Spelling error	B	x					x	
Special character	Q	x	x	x			x	
Word stem	Y							
Cropped	F	x	x	x			x	
Abbreviation	I			x				
Partially incorrect	J							
Plus/Minus	T	x	x	x	x	x	x	
Stop word	X							
Missing	E	x	x	x			x	
No author name	P							
Completely incorrect	D	x	x	x	x	x	x	
Space	K	x				x	x	Potential impact
Informational letter	L			x				
Additional information	N	x	x	x	x	x	x	
Punctuation	R				x	x		
Padded	S		x	x	x	x		
Full first name	U		x		x	x		
Interchanged fields	G					x		
Jumbled value	H			x			x	
Incorrect order of authors	O	x	x	x	x	x	x	
Incorrect interpretation of names	M		x				x	
Incorrect interpretation of additional information	V							
Typographical variation	A							
Spelling error	B		x	x	x	x		
Special character	Q							
Word stem	Y							
Cropped	F				x	x		
Abbreviation	I	x			x	x	x	
Partially incorrect	J	x			x	x	x	
Plus/Minus	T							
Stop word	X							
Missing	E				x	x		
No author name	P	x	x	x				
Completely incorrect	D							
Space	K		x	x	x			No impact
Informational letter	L	x	x		x	x	x	
Additional information	N							
Punctuation	R		x	x				
Padded	S							
Full first name	U			x				
Interchanged fields	G							
Jumbled value	H				x	x		
Incorrect order of authors	O							
Incorrect interpretation of names	M			x	x	x		
Incorrect interpretation of additional information	V	x	x	x	x	x	x	
Typographical variation	A	x	x	x	x	x	x	
Spelling error	B							
Special character	Q				x	x		
Word stem	Y	x	x	x	x	x	x	
Cropped	F							
Abbreviation	I		x					
Partially incorrect	J		x	x				
Plus/Minus	T							
Stop word	X	x	x	x	x	x	x	
Missing	E							
No author name	P				x	x	x	
Completely incorrect	D							
		WoS	Scopus	GS	CWTS	iFQ	SM	

Figure 31: Matrix of inaccuracies impacting the citation matching process

The data triangulation revealed that the IACs *D Completely incorrect* and *T Plus/Minus* have an *explicit impact* on the citation matching in all data sources. While *Completely incorrect* data values may be difficult to solve in the citation matching process, values which can be converted into the correct ones by calculation are not. The IACs *N Additional information* and *O Incorrect order of authors* have a *potential impact* on the citation matching in all data sources. *Additional information* in data values can be handled by matching rules which recognize that the entire, correct values are enclosed. The *Incorrect order of authors* is more difficult to surmount, since the cited reference information in WoS only contains the first author. Therefore, the incorrect values can only be counterbalanced by other (hopefully correct) data fields.

The IACs in the category *potential impact* do not necessarily impact the citation matching process. Since we do not exactly know how the matching algorithms work, some of the IACs occurring in the missed citations might have been handled by the algorithms if they had occurred individually, but the correct matching was hindered by a co-occurring IAC responsible for the non-match. Furthermore, IACs can also change their impact, depending on the bibliographic field and the variations which are tolerated in the citation matching algorithm. For example, the IAC *K Space* may be a minor inaccuracy in the field publication name if this field allows for space character variations, but it may lead to a non-match of an author's last name if this field only allows for exact matches (e.g. De Roover vs. DeRoover). Moreover, the IAC *V Incorrect interpretation of additional information* was identified as having *no impact* on the citation matching in our evaluation. The result could differ for a data sample where it occurs in the first author name and not in the second one. Hence, the classification of IACs should be regarded as an indication. Information about the weight each matching algorithm assigns to the different bibliographic data fields in the matching process would be required to derive less ambiguous conclusions.

8.4 Summary

While the domain, discipline and document type seem to impact non-matches of citations in WoS, other characteristics of publications, such as publication year and language of the cited and citing article, do not influence the occurrence of missed citations.

We compared five data sources and, implicitly, their respective citation matching algorithms with each other to determine whether they were able to match the missed WoS citations or

whether they were missed by these systems as well. In other words, are citations which WoS could not match to its cited articles also not matched in Scopus, GS and the databases of the three applied research groups? iFQ and CWTS were able to match the most missed citations, followed by Scopus and GS. Science-Metrix matched the fewest citations, which is most probably due to its matching algorithm usually employing the article title provided in the cited reference information by Scopus. GS, CWTS, iFQ and Science-Metrix did not cover between 2 and 4% missed WoS citations, whereas Scopus did not cover 12%.

The analysis revealed the types of inaccuracies the data sources are able to counterbalance in their citation matching algorithms and which inaccuracies lead to non-matched citations, i.e. lost or missed citations, which are, therefore, not considered in citation analyses. From our evaluation we conclude that IACs *D Completely incorrect* and *T Plus/Minus* represent the biggest concerns in the citation matching process, of which the IAC *T Plus/Minus* could be overcome by allowing for specific matching thresholds. In most cases *Omitted* data values (IAC *E*) lead to missed citations as well, although in the databases of CWTS and iFQ only the co-occurrence with another inaccurate field causes a non-match.

The results should not be an invitation to scientific authors to stop taking care in compiling their bibliographies, because algorithms can take care of the majority of incorrectly cited articles. Nevertheless, to err is human and, therefore, finding ways to compensate for inaccuracies caused by authors will remain an important part of citation analysis. Additionally, not all inaccuracies are caused by authors, but also by data handling procedures. The correction of these can either be directly tackled in the data ingestion or extraction process or together with other inaccuracies in the matching process.

9 Proposals to Improve the Process of Citation Matching

While only a small share of references is completely discrepancy-free, not all inaccuracies impact the citation matching process. The accuracy on a data value level is relatively high, especially in the fields that are used in the citation matching process. However, some citations are still missed by all matching algorithms. In this chapter, we explore different potential improvements to the citation matching process which originate from the results of chapters 6, 7 and 8. Since 85% of the data values of the references are accurate, we suggest that data fields other than those traditionally used in the citation matching process may be good candidates to compensate for inaccurate data in the customary fields. Even though this implies either using Scopus raw citation data or else an overdue change of data ingestion policy at WoS, we disregard these restrictions and make proposals for experiments on citation matching based on all assessed bibliographic fields. Section 9.1 and its subsections present proposals for each bibliographic field assessed in this research. They are complemented by sections 9.2 to 9.5, which discuss proposals specific to the different facets of the data sample, the formats of the data values (string vs. numerical fields) as well as the use of the DOI. Section 9.6 summarizes conclusions regarding the cited reference information provided by WoS, Scopus and GS. None of the proposals have been tested. Hence, their feasibility in terms of processing time as well as their actual impact on the precision of the citation matching will have to be investigated in future work (cf. section 10.2).

9.1 Bibliographic fields

From the results of our evaluation, we conclude that the publication year, the volume number and the starting page number should be given the greatest weight in the matching process, since they are the most accurate bibliographic fields. They should be followed by the first author-related fields, of which the last names and the first initial have a greater weight than the

second initial. Next in place come publication name, other author-related fields, followed by the ending page and finally the article title.

The following subsections discuss suggestions on how to rectify the inaccuracies detected in the different bibliographic fields.

9.1.1 Author-related fields

First of all, we propose removing punctuation and space characters from the author's last name in the matching process. Since the IAC *R Punctuation* has an explicit impact and the IAC *K Space* a potential impact on the citation matching of WoS and Science-Metrix, we conclude that they do not apply this kind of data parsing. Although, the IAC *U Full first name* also has an explicit impact on the WoS and Science-Metrix data, we refrain from making suggestions about this inaccuracy, since the cropping of the first name could hinder additional author disambiguation matching, which was not investigated in this thesis. For Scopus, CWTS and iFQ this inaccuracy is only of potential impact in the citation matching process, hence we cannot exclude the possibility that the non-match was caused by one or more co-occurring inaccuracies and their algorithms are quite able to handle a *Full first name*.

Spelling variations, such as *Spelling error*s (IAC *B*) and *Special character*s (IAC *Q*), and *Partially incorrect* (IAC *J*) as well as *Jumbled value*s (IAC *H*) can be overcome by fuzzy string matching methodologies allowing for variations in authors' last names. iFQ uses the Damerau-Levenshtein distance for author names with specific thresholds for reliability reasons and CWTS employs the soundex code[45] of last names if they cannot be uniquely matched (cf. Appendix A). *n*-grams could be another suitable option (Christen, 2006; cf. section 2.2.3). Hence, employing string matching methodologies in the matching process of author names is a necessity and needs to be tested out with adequate variation thresholds.

The *Incorrect interpretation of author names* (IAC *M*) has a potential impact in Scopus and Science-Metrix and an explicit impact in WoS. This implies that the other three data sources have found a way to correct this inaccuracy. On the one hand, this could be achieved by including different variations of author names from an authority file of names in the matching process. However, we know that neither iFQ nor CWTS uses additional resources (cf. Appendix A). Hence, the application of fuzzy string matching in their matching algorithms rectified this inaccuracy. Nevertheless, the integration of authority files or author registries,

[45] The soundex code is a phonetic algorithm that converts a string into a code according to its sound in the English language (Knuth, 1997).

such as ORCID[46] or Google Scholar Citations[47], are options that might not only support the citation matching, but also the author disambiguation process. Additionally, they might also provide a solution for *Interchanged* (IAC *G*), *Cropped* (IAC *F*), *Missing* (IAC *E Omitted*, IAC *P No author name*) or *Completely incorrect* (IAC *D*) first and second initials without allowing for too much variation in switching the fields in the matching process. Otherwise *Omitted* and *Completely incorrect* first and second initials can only be counterbalanced by confident matches of other bibliographic fields.

Even though the overall share of accurate data values is lower for authors other than the first ones (around 74% vs. 90%, respectively), we still advocate experimenting with additional authors in the citation matching process. Specifically for an *Incorrect order of authors* (IAC *O*) in references, additional author names could provide an opportunity to match citations correctly. While GS matched citations containing this particular inaccuracy correctly, no other data sources were able to handle it. However, it was not the sole reason for the non-match, but co-occurred with other inaccuracies. Therefore, we cannot ascertain that the incorrect order in fact caused the citations to be missed.

9.1.2 Article title

A study by Yannakoudakis et al. (1990) briefly discussed in section 2.2.3, as well as the information from Science-Metrix that they also usually employ the article title, has given rise to the proposal of experimenting with the article title in the citation matching process. Specifically in the SSH disciplines, the article title could be an alternative field to explore for citation matching, since, in the majority of cases, article titles are fairly accurate. Since more citing references use the original German article title than a translated English version, the original article title for articles in languages other than English would need to be available, i.e. not the translated version from WoS. Additionally, German articles were sometimes cited with *Completely incorrect* English translations of their titles. In this instance, Scopus provides the better raw citation data.

The use of the article title incurs more inaccuracies, such as the omission of subtitles (IAC *F Cropped*), the addition of non-existent subtitles (IAC *S Padded*) and *Typographical variation*s (IAC *A*). However, the majority of them can be overcome by appropriate string matching methodologies, such as the Sorted-Winkler function (cf. sections 2.2.3 and 9.4). Moreover, we

[46] Open Researcher and Contributor ID: https://orcid.org/
[47] http://scholar.google.com/intl/en/scholar/citations.html

found that a high score of the Levenshtein distance function can either be an indicator that the article title has been *Cropped* or that it is a translation of the original article title or that it is *Completely incorrect*. Therefore, the Levenshtein score could be used in the decision process of whether to use the article title in the citation matching process or not. To achieve the best results, the original article title as well as a translation, as already indexed and provided by Scopus, should be used.

Furthermore, for forthcoming publications, the article title could be an opportunity to compensate for an *Omitted* publication year, volume and page number.

9.1.3 Publication name

The publication name is a field in which a single inaccuracy did not lead to missed citations, but only a co-occurrence with other inaccurate fields. The results in section 8.3 also showed that the majority of data sources are able to handle the different *Abbreviation*s well, in particular Scopus. While we assume that fuzzy string matching is applied to match variations of publication names (similarly to author names) a central registry or authority file of publication names could prove to be of additional benefit for the matching. We conclude that the matching of this particular field already works relatively efficiently.

Nevertheless, we still suggest experimenting with the *Additional information* detected in publication names. For example, if the string *in press* is found and one or more of the numerical data fields (publication year, volume number, starting page) are *Omitted*, this could trigger the use of the article title in the matching process to compensate for the other *Omitted* data fields. The query *in press** in the field *Cited Work* of the *Cited Reference Search* in WoS retrieves almost 2.4 million records[48]. Even if only half of them are citations to articles actually indexed in WoS (because they could also be orphan references, i.e. citations to so-called non-source items, cf. Chi, 2013), this would be a valuable asset in remedying the number of non-matches in WoS.

9.1.4 Publication year

The inaccuracies related to *Added data values* in publication years are in most instances handled well by all data sources (IAC *L Informational letter*). With the exception of GS, of which we know least about the data extraction and matching process, it is of potential impact. A few *Omitted* publication years due to citations referring to forthcoming publication years

[48] Figure was last checked August 26, 2014.

might be counterbalanced by the matching of other data fields (cf. section 9.1.2 and 9.1.3). However, the problem of transposed publication years (IAC *T Plus/Minus*), e.g. the cited publication year is 1997 instead of 1998, could still be improved. In particular, iFQ's matching algorithm seems to be more conservative in this respect than that of CWTS. Yet, it is important to experiment with different thresholds to find an optimum balance between correct matches and false positives (cf. section 9.3). Again, *Omitted* (IAC *E*) or *Completely incorrect* (IAC *D*) publication years can only be compensated for by confident matches of other bibliographic fields.

9.1.5 Volume number

Inaccuracies caused by an issue number mistaken for the volume number (IAC *G Interchanged fields*) are best handled in the citation matching algorithm of iFQ. Only two instances occurred in the missed citations which were classified as of potential impact in the iFQ data. In all other data sources, this inaccuracy was of explicit impact on the non-matches. The full bibliographic information from the cited reference would provide an opportunity to detect a switch between issue and volume number. A rule could be applied which checks whether a switch of the two fields leads to a successful match. Since the issue number is not part of the cited reference information in WoS and only Scopus provides this data, we infer that the match of the volume number in the iFQ algorithm allows for a larger threshold than in the other algorithms or has less weight in the matching.

While *Omitted* (IAC *E*) and *Completely incorrect* (IAC *D*) volume numbers can again only be rectified by the correct matching of other bibliographic fields, transposed volume numbers (IAC *T Plus/Minus*) could be converted into the correct values. Analogously to the publication year, experiments would entail a suitable trade-off between correct matches and false positives. However, transposed volume numbers only occur in missed citations by WoS and Science-Metrix, thus, the other data sources seem to have already found the optimum threshold. Additionally, *Padded* volume numbers (IAC *S*) occur solely in the WoS and Science-Metrix data. The other data sources are apparently able to extract the correct value from a *Padded* volume number.

9.1.6 Pagination

In the case of starting page numbers, the most inaccuracies are caused by *Completely incorrect* data values (IAC *D*), which means that it will be almost impossible to find a data manipulation rule to convert incorrect into correct values. However, 51% of those in missed citations

actually held the cited page number. Hence, for these cases the use of the ending page number in the citation matching process could prove to be the solution. Furthermore, in addition to the other proposals discussed in the previous two sections, the use of the ending page for transposed (IAC *T Plus/Minus*), *Jumbled* (IAC *H*) or *Interchanged* (IAC *G*) page numbers would also provide an opportunity to correct a non-match of the starting page number.

The evaluation of the ending page number revealed that, when a reference gives an ending page number, it is completely accurate in about 60% of all cases. The inaccuracies to be dealt with are either *Abbreviated data values*, i.e. *Cropped* page numbers (IAC *F*), or transposed digits (IAC *T Plus/Minus*), or, in very few cases, additional values that do not belong to the correct pagination or any other numerical field (IAC *S Padded*). However, specifically for the identification of cited page numbers in references, the ending page numbers could be employed to define an interval between the starting and the ending page. If the page number cited in the reference matches the interval, the non-match with the starting page could be overruled and further successful matches with other bibliographic fields defined in the algorithm could still lead to a successful match.

9.2 Facet-specific proposals

Even though studies on inaccuracies in bibliometric data sources report on a highly skewed distribution of inaccuracies (e.g. Moed, 2005), we still found a few patterns that could be transformed into opportunities to increase the rate of successfully matched citations.

While the number of inaccuracies is higher in the NS than in the SSH, more missed citations occur in the SSH than in the NS. Hence, the inaccuracies occurring in the SSH impact the citation matching process more than those in the NS. The domain of the cited or citing article could, therefore, be employed to trigger domain-specific or even discipline-specific matching rules. Since citations in the SSH tend to follow the exact bibliographic data from the original article, this implies that allowance should be made for greater or different spelling variations of author names, in particular for publications in languages other than English, or for specific requirements if the article title is employed in the citation matching, i.e. the original as well as a translated version of the article title must be available. Due to the increased occurrence of references containing (only) the cited page number in the SSH, the above-mentioned rule for cited page numbers should be applied (cf. section 9.1.6) exclusively for references in the SSH. Since references in the NS tend to follow the bibliographic data of the WoS target records,

fewer variations in the matching may be required. The application of the article title in Chemistry is not recommended, since a high percentage of references do not cite it. Moreover, if the full bibliographic information is available, the use of additional authors will be limited to only a few more, since citing authors tend to use *et al.* more often. The domain could also play a role in the identification of false positives, since we found that 27% of the false positive source articles were not published in the same domain as their target articles. This finding is also evidence that WoS most probably does not use the domain information in its citation matching process.

The language of cited articles does not seem to influence missed citations, but references to English cited articles are more accurate than those to German ones. As described in section 7.5, this is due to the higher quantity of *Special character*s in German bibliographic data, as well as to the policy of WoS of indexing foreign-language articles solely with a translated English version. However, we also detected a tendency in references to English cited articles to use *et al.* more frequently and to omit the ending page. Hence, the tolerance for matching umlauts as well as the rule for including the ending page number to identify cited page numbers (cf. section 9.1.6) should only be applied to German target articles (and to those of other languages with frequent *Special character*s). Since German source articles tend to cite German article titles and publication names more accurately and to provide more complete citation information, whereas almost half of the English source articles, when citing a German target article, translate the article title into English, this must be considered in a rule if the article title is employed in the citation matching process.

Since the missed citation rates for Proceedings papers are higher than for other document types, we conclude that this specific document type may require additional matching rules. However, as reported in section 7.6, inaccuracies in all document types are relatively similarly distributed. Therefore, we are unable to make any specific suggestions for this document type and suggest further studies of a data sample exclusively for Proceedings papers.

9.3 Numerical data fields

In sections 9.1.4 to 9.1.6, we proposed that transposed numerical data values can still be correctly matched if the algorithm allows for an optimum variation threshold. The inaccuracies assessed as IAC *T Plus/Minus* denote a data value that is correct if the number 1 or 2 is added to, or subtracted from, the data value and we defined the calculation to be made either on the

total number or just on one of the digits. The IAC *H Jumbled value* was assigned to values where only the order of the single digits was jumbled and no additional calculation was necessary. Considering these two IACs, the relatively large threshold could introduce false positive matches as well. Since, Scopus, GS and CWTS seem to have found a way to compensate for the majority, but not all, of the transposed digits, it will be interesting to see if this result changes when all matched citations of the data sources, not only those missed by WoS, will be compared in future work. To rectify inaccuracies of the type *G Interchanged fields* this would, on the one hand, require the issue number in order to correct the majority of interchanged volume numbers (cf. section 9.1.5); on the other hand, a rule by which all numerical fields are iterated in the matching process could provide additional successful matches.

To improve the matching of numerical fields, experiments need to be conducted to find an optimum range of variation to optimize the precision and recall of the matching. Furthermore, this range could be expanded if combined with fairly stringent matching of other bibliographic data fields.

9.4 The use of string matching methodologies

String data, such as author-related fields, article title and publication name, are predestined fields for employing fuzzy string matching methodologies to rectify inaccuracies in their data values. Summarizing the application possibilities from sections 9.1.1 to 9.1.3, the IACs *A Typographical variation*s, *B Spelling error*, *F Cropped*, *H Jumbled value*, *J Partially incorrect*, *K Space*, *Q Special character*s, *S Padded* could be overcome by adequate variation thresholds in the matching process.

The incidences of the IAC *K Space*, as of potential impact on the citation matching process, speaks, on the one hand, for the elimination of all space characters in the data parsing process within one data field; on the other hand, the elimination of all space characters as well as the occurrences of the IACs *F Cropped*, *H Jumbled value*, *S Padded* and *J Partially incorrect* in string data values indicate the use of fuzzy string matching approaches, e.g. *n*-grams, Sorted-Winkler function or Bag distance, to identify whether the correct value is part of the data field to be matched. For instance, it can lead to a successful match of a *Jumbled* author name where the prefix or suffix is cited in a different location from that of the target article. Additionally, *Spelling variations* in author names and article titles can still be successfully matched if the

string matching function allows for an appropriate threshold. The applied bibliometric research groups CWTS and iFQ use string matching methodologies in their matching (cf. Appendix A). CWTS uses the soundex code for the last name if no unique match is achieved and, at some point, fuzzy string matching for publication names. iFQ uses the Damerau-Levenshtein distance for author names and publication names with thresholds for performance and reliability reasons. These solutions seem to work well and rectify the majority of inaccuracies, but could also be complemented by other approaches, in particular if the article title is included in the matching process.

9.5 The use of the DOI

iFQ stated to use the DOI in its citation matching, if available in the cited reference information of WoS (cf. Appendix A). The DOI is a digital identifier of an object that permanently identifies any physical, abstract or digital objects with a digital reference and is primarily used for documents (DOI, 2014). The DOI is provided by the publisher of the document and resolves through http://dx.doi.org/[plus the DOI] into the URL of the actual resource. The DOI itself is not a guarantee of accurate bibliographic data, since we learned during the assessment process[49] that publishers can also provide faulty records. However, it does guarantee the unique identification of an article and, therefore, facilitates the citation matching procedure.

If more and more publishers provided bibliographic records with a DOI to citation indexes like WoS and Scopus, the data cleaning and harmonization process of these well-linked documents would take up less and less of the database publishers' time. This would open up an opportunity to invest more resources in remedying already indexed records and resolve non-matched citations.

9.6 The cited reference information in WoS, Scopus and GS

In general, we conclude that neither WoS nor Scopus nor GS provide absolutely accurate citation counts and we support previous studies in stating that the choice of database(s) should

[49] For example, for the DOI 10.1080/00220270210134600 we found an incorrect article title, which in the meantime has been corrected by the publisher, Taylor & Francis. However, we also detected one example in our data which was matched to the correct cited article in spite of an incorrect publication year because the reference in the citing article contained the correct DOI.

depend on the goal of the bibliometric study (e.g. Meho & Yang, 2007). At all costs, the citation data should not be trusted blindly. The advantage of using WoS is the *Cited Reference Search*, which at least provides the possibility to search for citations the system was not able to match automatically, a feature lacking in Scopus and GS. However, the number of false positives identified in our WoS data sample signals that even the correct matches by WoS cannot be fully trusted. Since we found that, in some cases, neither the (broader) discipline nor even the domain matched the domain of the cited article, one should at least check for "suspicious" journal titles in the process of retrieving the citing articles. Determining how big the problem of false positive matches is in the other databases compared with WoS will be tackled in future work (cf. section 10.2).

In comparison to the other matching algorithms, the following problem areas of the WoS citation matching system were identified:

- Correct extraction of starting page numbers if a cited page number is given in addition.
- Matching a citation if it only gives the cited page number.
- Matching publication years that differ by only one or two years.
- Matching transposed digits in volume numbers and starting pages.
- Citations of forthcoming publications, where not only the publication name contains the addition *in press*, but also, in some cases, either one or a combination of the bibliographic fields publication year, volume number and starting page is missing.
- Distinguishing the volume number from the issue number if the citation style varies and verifying whether the volume and issue numbers in the citation might have been switched by the author. Example: *Heft 3, 54. Jg.* needs to be correctly identified as volume number 54 and issue number 3.
- Low tolerance of different spellings of special characters, such as Germanic umlauts. An umlaut is always converted to the letter without the umlaut and different spellings, such as *ue* or *ae*, are ignored and, therefore, not correctly matched.
- Punctuation in the first and second initial as well as citations giving the first given name in full instead of abbreviating it to the first initial. Even though these seem to be minor inaccuracies, we found examples where the only discrepancy in the CitRefmiss record was an additional dot after the first

initial and where the first full name was the only difference in the bibliographic data of a correctly matched citation.

Scopus's matching algorithm, specifically in the case of publication names, is less conservative than that of WoS. Additionally, the matching of *Interchanged fields* as well as jumbled or transposed values works better, but still causes citations to be missed in the process. Identified problem areas include:

- Records are more likely not to contain cited references at all.
- Records are more likely to omit certain references from the cited reference information. This problem was not specific to one discipline or domain, since one could argue that the data extraction process failed in the case of footnotes or endnotes in the SSH. It occurred in both domains and in different disciplines.
- Scopus also introduced additional inaccuracies into their data extraction process (e.g. *Incorrect interpretation of author names*).

The cited reference information in GS is non-existent for the user. Therefore, we based our evaluation on the assessment results of the original reference. However, until GS changes its policy of not providing an interface to download records and citation information, we cannot draw any definitive conclusions on what inaccuracies cause citations to be missed in GS.

Comparing the three citation indexes in terms of their suitability as a bibliometric data source for applying customized citation matching algorithms, we conclude that the cited reference information in Scopus, if present, is "paradise" for any citation matching algorithm, as it contains the citation data exactly as it is found in the original source article. Hence, it is possible to employ a variety of bibliographic fields in the citation matching. Even though Scopus does not use the article title in the matching process itself, its algorithm seems to have more discrepancy tolerance than that of WoS, which still needs to be corroborated in a future study comparing all matched and missed citations as well as false positive matches (cf. section 10.2). The occasionally missing cited reference information suggests that Scopus should be used as a complement to WoS in order not to miss citations simply because they were omitted from the cited reference list in the data extraction process. GS is not an adequate system for customized citation matching, since it does not provide easy access to cited reference information.

At the very start of the SCI, Eugene Garfield was already aware of the fact that indexing only five bibliographic fields per reference might not be sufficient (Garfield, 1990). At that time the decision was a financial one. However, in times of cheap mass storage and extensive possibilities of data matching, the reasons why Thomson Reuters has not changed its policy of citation indexing is not transparent. The argument that the database was first and foremost invented as a literature database is, 50 years later and considering all the technological progress, no longer convincing. Even if Thomson Reuters argues that citation data is not its core business, it does provide citation counts in its system and, by offering a product, it should be of good quality (Tenopir, 1995). The longer the change is postponed, the more publications will be matched according to outdated matching procedures. In conclusion, we advocate a change in the ingestion procedure of cited reference information in WoS.

9.7 Summary

Apart from the *Completely incorrect* publication years, volume numbers and starting pages that can only be counterbalanced by correct matching of other fields, there are inaccuracies that could be handled by allowing for thresholds in the matching process. In general, we suggest including as many bibliographic fields as possible in the citation matching. The more fields are involved, the more flexibility this will give the citation matching algorithm. Even though one might argue that the matching of the five fields coupled with occasional use of the DOI works well for approximately 90% of the data in WoS, the fact that we not only found non-matched citations, but also false positive matches in our relatively small data sample cannot be ignored. It shows that we do not know how many of the 90% we can really trust. In spite of the fact that the false positives only account for 1%, we do not know anything about their distribution in WoS. Additionally, the mere fact that we found false positives, even though WoS's citation matching is so conservative, points to the fact that a matching algorithm does not need to be conservative, but needs to find a maximum of unique correct matches and a minimum of false positive matches. Moreover, we speculate that, at the time when the policy of matching citations with the five bibliographic fields was established by WoS, no experiments with other bibliographic fields were conducted. Hence, it is impossible for WoS to actually have found the best possible trade-off between false positives and missed citations.

None of the proposals made in this chapter have been tested. Hence, the feasibility in terms of processing time as well as actual impact on the precision of the citation matching will have to be investigated in future work. A trade-off between correct and false positive matches needs to

be found, whereby the goal must be an optimum balance between recall and precision. Additionally, the system's performance as well as invested resources must be in relation to the achieved outcome, since it is likely that, for the last 10% of data, one would need hundreds of rules, almost working on a case-by-case basis (Dasu & Johnson, 2003).

10 Conclusion

This doctoral research investigates data accuracy in bibliometric data sources and its impact on citation matching. We defined inaccuracies in bibliometric data sources as discrepancies in data values in bibliographic references, since they are the essential part of the citation matching process and, therefore, have the greatest impact on its accuracy. A data sample, consisting of typical cases of publications in WoS, was assessed to identify prevailing inaccuracies in bibliographic references which can interfere with the citation matching process. In a qualitative content analysis, inaccuracies were examined, categorized and summarized into a taxonomy of bibliographic inaccuracies. To determine which of these inaccuracies in fact influence the citation matching process, a specific subset, i.e. missed citations in WoS, was investigated and triangulated with five other data sources. This chapter summarizes the results and contribution of this research in section 10.1. Section 10.2 discusses future work.

10.1 Contribution

The main contribution of this dissertation is the systematic investigation of inaccuracies in citations. By choosing a stratified purposeful sample, we examined a sub-universe of the entire WoS data universe, which allows a generalization of the overall findings. We did not aim to estimate an overall error rate for the different bibliometric data sources, but drew conclusions from the patterns of inaccuracies in the different facets of the stratified data sample as well as from the results of the evaluation of missed citations. The three research questions addressed were:

RQ1 What types of inaccuracies occur in bibliographic data?
- How can they be categorized?
- How frequent is their incidence in bibliometric data sources?
- Can patterns be identified?

RQ2 What types of inaccuracies cause missed citations?
- How well do citation matching algorithms handle inaccurate data?

RQ3 How can the number of non-matches in the citation matching process be reduced?

To answer RQ1, we investigated the inaccuracies occurring in bibliographic references of a broad data sample in a qualitative content analysis (cf. chapter 6). The data sample was selected purposefully, representing typical cases of publications in WoS as well as different coverage facets (e.g. publication year, language, etc.). Since an inaccuracy was defined as any discrepancy between the value in the reference and the value defined as correct (from the original article or the WoS record), the inaccuracies not only describe the negligence of authors when compiling their bibliographies, but also capture different citation styles as well as specifics of the data structure in WoS. Therefore, the number of inaccuracies in a bibliographic reference is not an indicator of the increased possibility of being missed in the citation matching process. 15% of all references did not contain a single discrepancy, whereas 85% of all assessed data values (from all references) were discrepancy-free. The inaccuracies were categorized according to common characteristics (*Added data values*, *Disarranged data values*, *Incorrect interpretation of data values*, *Spelling variations*, *Abbreviated data values*, *Other variations*, *Not assessable*, *Missing data values* and *Completely incorrect*) and then grouped into a taxonomy according to the sophistication required of a rule which would convert a discrepant value into the correct one (*simple*, *moderate*, *complex*). This taxonomy is the first of its kind to describe bibliographic inaccuracies and can be used in the future for assessing the data accuracy of citation indexes with the goal of measuring error rates.

Chapter 7 summarizes the quantitative results of inaccuracies in bibliographic references. The most frequent inaccuracies are related to *Abbreviated* and *Missing data values*. The most accurate bibliographic fields are publication year, volume number and starting page; the least accurate is the article title. 90% of all data values are accurate in the WoS target records. Even though three of the most important matching fields, publication years, volume numbers and starting pages, are completely accurate in WoS, in our data sample there is still a 10% chance of encountering inaccurate target values in the citation matching process. References in the SSH reproduce citations more accurately according to the original bibliographic data, whereas references in the NS tend to shorten them. *Other variations* and *Completely incorrect* data values occur more often in the SSH, whereas *Missing data values* are more common in the NS. References to German target articles as well as references in German citing articles reflect the characteristics of the German language and, therefore, the citations also tend to follow the exact bibliographic data in the original article. No specific patterns were identified for different document types. However, since the number of references was relatively limited for

document types other than articles and reviews, this result may be refuted in a subsequent analysis with a larger data sample. The number of inaccuracies decreases over time, which suggests that citing authors are now compiling their bibliographies with more care and/or that the matching algorithms have improved. While *Disarranged* and *Missing data values* decrease, *Completely incorrect* and *Not assessable* data values increase over time.

Answering RQ2, we specifically assessed the cited reference information of missed citations in WoS, which were identified in the *Cited Reference Search*, against the cited reference information of correctly matched citations (cf. chapter 8). The WoS citation matching algorithm seems to be very conservative and does not appear to allow for many variations, except for deviating second initials and different *Abbreviation*s of publication names. In total, 57% of inaccuracies still present in the cited reference information of missed citations were caused by authors, 12% were due to the citation style which WoS was not able to process correctly, and 31% of inaccuracies were traced back to inaccuracies introduced in the data handling process or to data which had already been inaccurately provided to WoS.

The comparison with the other five data sources corroborates the fact that applied bibliometric research groups have developed sophisticated matching algorithms to match citations better. However, this result still needs to be verified by a full comparison of matched and missed citations as well as false positive matches. Considering the types and number of inaccuracies in the cited reference information of WoS, the algorithms of CWTS and iFQ work really well. The Science-Metrix algorithm does not match significantly more citations than the WoS algorithm, which is most probably due to the fact that they also usually incorporate Scopus's cited reference information and, by that more bibliographic data, i.e. the article title, into their matching process. Even though Scopus seems not to make use of more bibliographic fields in its citation matching, it also matches a considerable number of missed WoS citations. GS provides the largest coverage of missed WoS citations and also matches over 60% of missed WoS citations. Since we did not perform an overall comparison of overlap and coverage of all citations, the results need to be corroborated in future work.

The data triangulation revealed that *Completely incorrect* starting page numbers and transposed publication years may cause a citation to be missed in all data sources. However, it is more often a combination of more than one kind of inaccuracy in more than one field that leads to a non-match. Comparing all algorithms, GS specifically handles an *Incorrect order of authors* well, while Scopus' algorithm is able to match the widest variety of *Abbreviation*s of publication names. The algorithms of CWTS and iFQ show the best results for matching

Special character variations as well as *Jumbled value*s. Furthermore, iFQ manages *Interchanged fields* the best. The Science-Metrix algorithm only handled the *Incorrect interpretation of author names* better than WoS. However, these inaccuracies were also rectified by GS, CWTS and iFQ. In conclusion, CWTS can be trusted most with respect to handling inaccurate data, since iFQ's algorithm works slightly better, but is not in production yet. Science-Metrix's algorithm would need to be tested with Scopus data to derive more conclusive results.

RQ3 is answered in chapter 9. It proposes rules and changes that could be applied in the citation matching process. Even though they are all non-tested, they comprise a set of unpublished proposals to further enhance citation matching processes in bibliometric data sources, based on the data and analysis of the dissertation. The approaches encompass all bibliographic fields assessed in the qualitative content analysis as well as the DOI. Furthermore, facet-specific rules were derived from the quantitative analysis of inaccuracies and a summary of how inaccuracies in string and numerical data fields could be rectified is given. We conclude by addressing a bold suggestion to Thomson Reuters to change their cited reference ingestion policy, while also discussing the complementary use of Scopus and not recommending the use of GS for customized citation matching.

The results of this dissertation contribute to the research on increasing the transparency of citation analysis results. With the systematic investigation of inaccuracies in citations, we provide an improved understanding of inaccuracies in bibliographic data. We have laid the foundation for bibliographic data accuracy assessment which can measure error rates in bibliometric data sources. Moreover, this research presents unique findings by comparing and analyzing the non-published citation matching algorithms of three leading applied bibliometric research groups, CWTS, iFQ and Science-Metrix, without infringing their competitive advantage, with the citation matching results of the three main citation indexes used for bibliometric analyses, WoS, Scopus and GS. Hence, the results, in particular the proposals formulated in chapter 9, could trigger changes in the citation matching procedures of all data sources. Moreover, the applied bibliometric research groups can benefit from the gold standard created in this research, i.e. the data corpus consisting of manually checked citations, providing them with an ideal reference point to further experiment with citation data and their matching algorithms.

10.2 Future Work

First of all, future work will entail an investigation of all data sources according to matched and missed citations as well as false positives for our data sample (cf. limitations in section 8.2). We will examine whether all matched citations in WoS were matched in the other data sources as well and whether the other data sources also include (the same) false positive matches (as WoS). A comparison with the inaccuracies that cause differences in this present evaluation will contribute to further pinpointing rules that can be experimented with in citation matching algorithms. Moreover, it will be interesting to compare the occurrences of inaccuracies in matched references. For example, we found a reference with an incorrect publication year (IAC *T Plus/Minus*), which was correctly matched by WoS because of the DOI that was cited in the references. Hence, to further improve our understanding of citation matching it is necessary to determine what kinds of inaccuracies are compensated for by other bibliographic fields.

Secondly, the proposals presented in chapter 9 will need to be tested for their effectiveness in the citation matching procedure of the different data sources as well as in terms of system's performance and invested resources. This could be achieved in cooperation with the applied bibliometric research groups and using the corpus of manually verified citations compiled in the present research.

The results of the distribution of inaccuracies in the facet *document type* did not reveal any specific patterns. However, the missed citation rate of some document types, e.g. Proceedings papers, was higher than for others. Therefore, future work should focus on a detailed investigation of references in document types other than Articles. This will establish whether references in some document types simply contain more inaccuracies than others or whether it is possible to pinpoint types of inaccuracies that lead to missed citations.

Furthermore, future work could include studying the influence of self-citation on the accuracy of references. In our data sample, 12% of the missed citations were caused by authors who cited their own work. It would be beneficial to investigate whether authors citing their own work tend to cite it correctly or whether the references are as error-prone as those by other authors. The results could lead to an additional rule to consider in the citation matching process.

The taxonomy developed to describe and categorize bibliographic inaccuracies could be expanded in future work and could also be drawn on to assess the data accuracy of citation indexes. This may result in metrics that could be used to determine which data source to involve for which bibliometric study. For instance, it would be valuable to weight inaccuracies and bibliographic fields in different calculations, since, for macro-level studies, specific inaccuracies may be less serious than when only studying a research unit. In this respect, it would also be interesting to study whether data quality metrics correlate with missed citations, and whether missed citations can be identified as outliers in the database.

This dissertation has contributed to the field of bibliometrics and the question of how accurate the results of citation analyses are. The results of this research provide an improved understanding of the inaccuracies occurring in bibliometric data sources and demonstrate what inaccuracies persist and thus interfere with the citation matching process. It represents a first step towards making citation analyses and their matching processes more transparent. Only the transparency of all resources and instruments employed can ensure accurate results and, therefore, confidence in the results of citation analysis. We support the findings of previous studies that, given a correctly employed methodology coupled with complete and accurate citation data, citation analysis can be a competent tool to support research evaluation.

References

Oxford English Dictionary: Online version (3rd ed.) (2013). Retrieved September 24, 2014 from http://www.oed.com/

Abdulhayoglu, M. A., & Thijs, B. (2013). Matching bibliographic data from publication lists with large databases using n-grams. In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. F. Moed (Eds.), *Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics* (pp. 1151–1158). Vienna, Austria. Retrieved September 24, 2014 from http://www.issi2013.org/proceedings.html

Adriaanse, L. S., & Rensleigh, C. (2013). Web of Science, Scopus and Google Scholar. A content comprehensiveness comparison. *The Electronic Library, 31*(6), 727–744. http://dx.doi.org/10.1108/el-12-2011-0174

Aksnes, D. W. (2008). When different persons have an identical author name. How frequent are homonyms? *Journal of the American Society for Information Science and Technology, 59*(5), 838–841. http://dx.doi.org/10.1002/asi.20788

Archambault, É., Campbell, D., Gingras, Y., & Larivière, V. (2009). Comparing bibliometric statistics obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology, 60*(7), 1320–1326. http://dx.doi.org/10.1002/asi.21062

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries, 3*, Article no. 7. http://dx.doi.org/10.1186/1742-5581-3-7

Ball, R., & Tunger, D. (2006). Science indicators revisited – Science Citation Index versus Scopus: A bibliometric comparison of both citation databases. *Information Services and Use, 26*(4), 293–301. Retrieved September 24, 2014 from http://iospress.metapress.com/content/9VY6XD722HU17RBC

Ballou, D. P., & Pazer, H. L. (1985). Modeling data and process quality in multi-input, multi-output information systems. *Management Science, 31*(2), 150–162. http://dx.doi.org/10.1287/mnsc.31.2.150

Bar-Ilan, J. (2006). An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes. *Information Processing & Management, 42*(6), 1553–1566. http://dx.doi.org/10.1016/j.ipm.2006.03.019

Bar-Ilan, J. (2008). Which *h*-index? A comparison of WoS, Scopus and Google Scholar. *Scientometrics, 74*(2), 257–271. http://dx.doi.org/10.1007/s11192-008-0216-y

Bar-Ilan, J. (2010). Citations to the "Introduction to informetrics" indexed by WOS, Scopus and Google Scholar. *Scientometrics, 82*(3), 495–506. http://dx.doi.org/10.1007/s11192-010-0185-9

Batini, C., Cabitza, F., Cappiello, C., & Francalanci, C. (2008). A comprehensive data quality methodology for web and structured data. *International Journal of Innovative Computing and Applications, 1*(3), 205–218. http://dx.doi.org/10.1504/ijica.2008.019688

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*(3), 16:1-16:52. http://dx.doi.org/10.1145/1541880.1541883

Batini, C., & Pernici, B. (2006). Data quality management and evolution of information systems. In D. Avison, S. Elliot, J. Krogstie, & J. Pries-Heje (Eds.), *IFIP International Federation for Information Processing. The Past and Future of Information Systems: 1976–2006 and Beyond* (pp. 51–62). Springer US. http://dx.doi.org/10.1007/978-0-387-34732-5_5

Batini, C., & Scannapieco, M. (2006). *Data quality: Concepts, methodologies and techniques.* Berlin, New York: Springer.

Bauer, K., & Bakkalbasi, N. (2005). An examination of citation counts in a new scholarly communication environment. *D-Lib Magazine, 11*(9). http://dx.doi.org/10.1045/september2005-bauer

Beck, S. E., & Manuel, K. (2008). *Practical research methods for librarians and information professionals.* New York: Neal-Schuman Publishers.

Belew, R. K. (2005). Scientific impact quantity and quality: Analysis of two sources of bibliographic data. Retrieved June 17, 2012 from http://arxiv.org/abs/cs/0504036

Bennett, D. B., & Williams, P. (2006). Name authority challenges for indexing and abstracting databases. *Evidence based library and information practice, 1*(1), 37–57. Retrieved September 24, 2014 from http://ejournals.library.ualberta.ca/index.php/EBLIP/article/viewArticle/7

Bergstrom, C. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News, 68*(5), 314–316. Retrieved September 24, 2014, http://crln.acrl.org/content/68/5/314.short

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology, 55*(14), 1216–1227. http://dx.doi.org/10.1002/asi.20077

Bornmann, L., & Marx, W. (2013). How good is research really? *EMBO Reports, 14*(3), 226–230. http://dx.doi.org/10.1038/embor.2013.9

Bornmann, L., Mutz, R., Neuhaus, C., & Daniel, H. D. (2008). Citation counts for research evaluation: Standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics, 8*(1), 93–102. http://dx.doi.org/ 10.3354/esep00084

Bovee, M., Srivastava, R. P., & Mak, B. (2003). A conceptual framework and belief-function approach to assessing overall information quality. *International Journal of Intelligent Systems, 18*(1), 51–74. http://dx.doi.org/10.1002/int.10074

Braun, T., Glänzel, W., & Schubert, A. (1985). *Scientometric indicators: A 32 country comparative evaluation of publishing performance and citation impact.* Singapore, Philadelphia: World Scientific.

Buchanan, R. A. (2006). Accuracy of cited references: The role of citation databases. *College & Research Libraries, 67*(4), 292-303. http://dx.doi.org/10.5860/crl.67.4.292

Cameron, B. D. (2005). Trends in the usage of ISI bibliometric data: Uses, abuses, and implications. *Libraries and the Academy, 5*(1), 105–125. http://dx.doi.org/10.1353/pla.2005.0003

Chang, C.-L., McAleer, M., & Oxley, L. (2011). Great expectatrics: Great papers, great journals, great econometrics. *Econometric Reviews, 30*(6), 583–619. http://dx.doi.org/10.1080/07474938.2011.586614

Chi, P.-S. (2013). Do non-source items make a difference in the social sciences? In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. F. Moed (Eds.), *Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics* (pp. 612–625). Vienna, Austria. Retrieved September 24, 2014 from http://www.issi2013.org/proceedings.html

Christen, P. (2006). A comparison of personal name matching: Techniques and practical Issues. In S. Tsumoto, C.W. Clifton, N. Zhong, X. Wu, J. Liu, B.W. Wah & Y.-M. Cheung

(Eds.), *Sixth IEEE International Conference on Data Mining - Workshops (ICDMW)* (pp. 290-294). Hong Kong. http://dx.doi.org/10.1109/ICDMW.2006.2

Companjen, B. (2013). *Probabilistically matching author names to researchers.* Master thesis, University of Twente. Retrieved September 24, 2014 from http://essay.utwente.nl/63407/

Costas, R., & Bordons, M. (2008). Is *g*-index better than *h*-index? An exploratory study at the individual level. *Scientometrics, 77*(2), 267–288. http://dx.doi.org/10.1007/s11192-007-1997-0

Croneis, K. S., & Henderson, P. (2002). Electronic and digital librarian positions: A content analysis of announcements from 1990 through 2000. *The Journal of Academic Librarianship, 28*(4), 232–237. http://dx.doi.org/10.1016/s0099-1333(02)00287-2

Cronin, B. (1982). *The education of library-information professionals: A conflict of objectives? Aslib occasional publication: no. 28.* London: Aslib.

CWTS (2014). *Leiden University Ranking,* Retrieved September 24, 2014 from http://www.leidenranking.com/

D'Angelo, C. A., Giuffrida, C., & Abramo, G. (2011). A heuristic approach to author name disambiguation in bibliometrics databases for large-scale research assessments. *Journal of the American Society for Information Science and Technology, 62*(2), 257–269. http://dx.doi.org/10.1002/asi.21460

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning.* New York: Wiley-Interscience.

Davis, P. M. (2008). Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts? *Journal of the American Society for Information Science and Technology, 59*(13), 2186–2188. http://dx.doi.org/10.1002/asi.20943

De, S., Jones, T., Brazier, H., Jones, A. S., & Fenton, J. E. (2001). The accuracy of MEDLINE and journal contents pages for papers published in clinical otolaryngology. *Clinical Otolaryngology and Allied Sciences, 26*(1), 39–42. http://dx.doi.org/10.1046/j.1365-2273.2001.00414.x

Denzin, N. K. (1989). *The research act: A theoretical introduction to sociological methods* (3rd ed.). Englewood Cliffs, N.J.: Prentice Hall.

Dinkel, W. P. (2011). How do matchkeys affect citation counts? First steps towards an error calculus for bibliometric indicators. In E. Noyons, P. Ngulube, & J. Leta (Eds.), *Proceedings of the ISSI 2011 conference. 13th International Conference of the International Society for Scientometrics & Informetrics* (pp.175–180). Durban, South Africa.

DOI (2014). Retrieved September 24, 2014 from http://dx.doi.org

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics. Library and information science.* Amsterdam, New York: Elsevier/Academic Press.

Egghe, L. (2006). Theory and practise of the *g*-index. *Scientometrics, 69*(1), 131–152. http://dx.doi.org/10.1007/s11192-006-0144-7

Elsevier B.V. (2014a). *Elsevier announces launch of program to include cited references for archival content in Scopus.* Retrieved April 19, 2014 from http://www.elsevier.com/about/press-releases/science-and-technology/elsevier-announces-launch-of-program-to-include-cited-references-for-archival-content-in-scopus

Elsevier B.V. (2014b). *Finding accented and special characters.* Retrieved September 10, 2014 from http://help.scopus.com/Content/h_specialchars.htm

Elsevier B.V. (2014c). *Scopus: Content Overview.* Retrieved January 12, 2014 from http://www.elsevier.com/online-tools/scopus/content-overview

Even, A., & Shankaranarayanan, G. (2007). Utility-driven assessment of data quality. *SIGMIS Database*, *38*(2), 75–93. http://dx.doi.org/ 10.1145/1240616.1240623

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal, 22*(2), 338–342. http://dx.doi.org/10.1096/fj.07-9492lsf

Franceschet, M. (2010a). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics, 83*(1), 243-258. http://dx.doi.org/10.1007/s11192-009-0021-2

Franceschet, M. (2010b). Ten good reasons to use the Eigenfactor™ metrics. *Information Processing & Management, 46*(5), 555–558. http://dx.doi.org/10.1016/j.ipm.2010.01.001

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2013a). A novel approach for estimating the omitted-citation rate of bibliometric databases with an application to the field of bibliometrics. *Journal of the American Society for Information Science and Technology, 64*(10), 2149–2156. http://dx.doi.org/10.1002/asi.22898

Franceschini, F., Maisano, D., & Mastrogiacomo, L. (2013b). Research quality evaluation: comparing citation counts considering bibliometric database errors. *Quality & Quantity*, 1–11. http://dx.doi.org/10.1007/s11135-013-9979-1

Franceschini, F., & Maisano, D. (2011). Influence of database mistakes on journal citation analysis: remarks on the paper by Franceschini and Maisano, QREI (2010). *Quality and Reliability Engineering International, 27*(7), 969–976. http://dx.doi.org/10.1002/qre.1174

Frandsen, T. F., & Nicolaisen, J. (2008). Intradisciplinary differences in database coverage and the consequences for bibliometric research. *Journal of the American Society for Information Science and Technology, 59*(10), 1570–1581. http://dx.doi.org/10.1002/asi.20817

Galvez, C., & de Moya-Anegón, F. (2006). The unification of institutional addresses applying parametrized finite-state graphs (P-FSG). *Scientometrics, 69*(2), 323–345. http://dx.doi.org/10.1007/s11192-006-0156-3

Galvez, C., & de Moya-Anegón, F. (2007). Standardizing formats of corporate source data. *Scientometrics, 70*(1), 3–26. http://dx.doi.org/10.1007/s11192-007-0101-0

García-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of *h*-indices in psychology. *Journal of the American Society for Information Science and Technology, 61*(10), 2070–2085. http://dx.doi.org/10.1002/asi.21372

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science, 178*(4060), 471–479. http://dx.doi.org/10.1126/science.178.4060.471

Garfield, E. (1981). What's in a surname? *Naturwissenschaften, 68*(10), 519–520. http://dx.doi.org/10.1007/bf00365376

Garfield, E. (1983). Quality control at ISI: A piece of your mind can help us in our quest for error-free bibliographic information. *Essays of an Information Scientist, 6*, 144–151. Retrieved September 24, 2014 from http://garfield.library.upenn.edu/volume6.html

Garfield, E. (1990). Journal editors awaken to the impact of citation errors. *Essays of an Information Scientist, 13*(41), 367. Retrieved September 24, 2014 from http://garfield.library.upenn.edu/volume13.html

Garfield, E. (2005). Notes by Eugene Garfield (chapter 12.2). In H. F. Moed (Ed.), *Citation Analysis in Research Evaluation, Information Science and Knowledge Management: Vol. 9.* (pp. 170–172). Dordrecht: Springer.

Gibbs, G. (2007). *Analyzing qualitative data. SAGE qualitative research kit.* Los Angeles: Sage Publications.

Glänzel, W. (1996). The need for standards in bibliometric research and technology. *Scientometrics, 35*(2), 167–176. http://dx.doi.org/10.1007/BF02018475

Goldberg, R., Newton, E., Cameron, J., Jacobson, R., Chan, L., Bukata, W. R., & Rakab, A. (1993). Reference accuracy in the emergency medicine literature. *Annals of Emergency Medicine, 22*(9), 1450–1454. http://dx.doi.org/10.1016/S0196-0644(05)81995-X

Haas, S. W., & Grams, E. S. (2000). Readers, authors, and page structure: A discussion of four questions arising from a content analysis of web pages. *Journal of the American Society for Information Science, 51*(2), 181–192. http://dx.doi.org/ 10.1002/(SICI)1097-4571(2000)51:2<181::AID-ASI9>3.0.CO;2-8

Han, H., Giles, L., Zha, H., Li, C., & Tsioutsiouliklis, K. (2004). Two supervised learning approaches for name disambiguation in author citations. In H. Chen, H. D. Wactlar, C.-C. Chen, E.-P. Lim & M. G. Christel (Eds.), *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2004* (pp. 296–305). Tucson, AZ, USA. http://dx.doi.org/10.1109/JCDL.2004.1336139

Harzing, A.-W. (2008). *Google Scholar - a new data source for citation analysis.* Retrieved April 19, 2014 from http://www.harzing.com/pop_gs.htm

Harzing, A.-W. (2013a). A preliminary test of Google Scholar as a source for citation data: a longitudinal study of Nobel prize winners. *Scientometrics, 94*(3), 1057–1075. http://dx.doi.org/10.1007/s11192-012-0777-7

Harzing, A.-W. (2013b). Document categories in the ISI Web of Knowledge: Misunderstanding the social sciences? *Scientometrics, 94*(1), 23–34. http://dx.doi.org/10.1007/s11192-012-0738-1

Harzing, A.-W., & van der Wal, R. (2009). A Google Scholar *h*-index for journals: An alternative metric to measure journal impact in economics and business. *Journal of the American Society for Information Science and Technology, 60*(1), 41–46. http://dx.doi.org/10.1002/asi.20953

Havemann, F. (2009). *Einführung in die Bibliometrie.* Berlin: Gesellschaft für Wissenschaftsforschung e.V.; Institut für Bibliotheks- und Informationswissenschaft der Humboldt-Universität zu Berlin. Retrieved September 24, 2014 from http://edoc.hu-berlin.de/docviews/abstract.php?id=29712

Henzinger, M., Suñol, J., & Weber, I. (2010). The stability of the *h*-index. *Scientometrics, 84*(2), 465–479. http://dx.doi.org/10.1007/s11192-009-0098-7

Hildebrandt, A. L., & Larsen, B. (2008). *Reference and citation errors: A study of three law journals*. Presented at the 13th Nordic Workshop on Bibliometrics and Research Policy. 11-12 September 2008, Tampere, Finland.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569–16572. http://dx.doi.org/10.1073/pnas.0507655102

Hood, W. W., & Wilson, C. S. (2003). Informetric studies using databases: Opportunities and challenges. *Scientometrics, 58*(3), 587–608. http://dx.doi.org/10.1023/B:SCIE.0000006882.47115.c6

Huang, J., Ertekin, S., & Giles, C. L. (2006). Efficient name disambiguation for large-scale databases. In J. Fürnkranz, T. Scheffer, & M. Spiliopoulou (Eds.), *Knowledge Discovery in Databases: PKDD 2006 (Lecture Notes in Computer Science)* (pp.536–544). Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/11871637_53

IFLA (1998). *Functional requirements for bibliographic records.* Retrieved September 10, 2014 from http://www.ifla.org/files/assets/cataloguing/frbr/frbr.pdf

International Organization for Standardization (ISO) (2005). ISO 9000:2005: Quality management systems – Fundamentals and vocabulary. Geneva.

Jaćimović, J., Petrović, R., & Živković, S. (2010). A citation analysis of Serbian Dental Journal using Web of Science, Scopus and Google Scholar. *Serbian Dental Journal, 57*(4), 201–211. http://dx.doi.org/10.2298/sgs1004201j

Jacsó, P. (1995). Testing the quality of CD-ROM databases. In R. Basch (Ed.), *Electronic information delivery. Ensuring quality and value* (pp. 141–168). Aldershot, Hampshire, England, Brookfield, Vt., USA: Gower.

Jacsó, P. (1997). Content evaluation of databases. *Annual Review of Information Science and Technology (ARIST), 32*, 231–267.

Jacsó, P. (2004). The future of citation indexing: An interview with Eugene Garfield. *Online, 28*(1), 38–40.

Jacsó, P. (2005a). As we may search. Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science, 89*(9), 1537-1547. Retrieved September 24, 2014 from http://choo.fis.utoronto.ca/FIS/courses/LIS1325/Readings/jacso.pdf

Jacsó, P. (2005b). Comparison and analysis of the citedness scores in Web of Science and Google Scholar. In E. A. Fox, E. J. Neuhold, P. Premsmit, & V. Wuwongse (Eds.), *Digital Libraries: Implementing Strategies and Sharing Experiences. Proceedings of 8th International Conference on Asian Digital Libraries, ICADL 2005 (Lecture Notes in Computer Science: Vol. 3815)* (pp. 360–369). Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/11599517_41

Jacsó, P. (2005c). Google Scholar: The pros and the cons. *Online Information Review, 29*(2), 208–214. http://dx.doi.org/10.1108/14684520510598066

Jacsó, P. (2006). Deflated, inflated and phantom citation counts. *Online Information Review, 30*(3), 297–309. http://dx.doi.org/10.1108/14684520610675816

Jacsó, P. (2008a). The plausibility of computing the *h*-index of scholarly productivity and impact using reference-enhanced databases. *Online Information Review, 32*(2), 266–283. http://dx.doi.org/10.1108/14684520810879872

Jacsó, P. (2008b). The pros and cons of computing the *h*-index using Google Scholar. *Online Information Review, 32*(3), 437–452. http://dx.doi.org/10.1108/14684520810889718

Jacsó, P. (2008c). The pros and cons of computing the *h*-index using Scopus. *Online Information Review, 32*(4), 524–535. http://dx.doi.org/10.1108/14684520810897403

Jacsó, P. (2008d). The pros and cons of computing the *h*-index using Web of Science. *Online Information Review, 32*(5), 673–688. http://dx.doi.org/10.1108/14684520810914043

Jacsó, P. (2009). The *h*-index for countries in Web of Science and Scopus. *Online Information Review, 33*(4), 831–837. http://dx.doi.org/10.1108/14684520910985756

Jarke, M., Lenzerini, M., Vassiliou, Y., & Vassiliadis, P. (2003). *Fundamentals of data warehouses* (2nd ed.). Berlin, Heidelberg, New York: Springer. http://dx.doi.org/10.1007/978-3-662-05153-5

Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: product and service performance. *Communications of the ACM, 45*(4), 184–192. http://dx.doi.org/10.1145/505248.506007

Katz, J. S. (1999). The self-similar science system. *Research Policy, 28*(5), 501–517. http://dx.doi.org/10.1016/S0048-7333(99)00010-4

Knuth, D. E. (1997). *The art of computer programming Volume 3.* Reading, Massachusetts: Addison-Wesley.

Kostoff, R. N. (2002). Citation analysis of research performer quality. *Scientometrics, 53*(1), 49–71. http://dx.doi.org/10.1023/A:1014831920172

Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology, 58*(7), 1055–1065. http://dx.doi.org/10.1002/asi.20584

Kousha, K., & Thelwall, M. (2008). Sources of Google Scholar citations outside the Science Citation Index: A comparison between four science disciplines. *Scientometrics, 74*(2), 273–294. http://dx.doi.org/10.1007/s11192-008-0217-x

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, California: Sage Publications.

Larsen, B., Hytteballe Ibanez, K., & Bolling, P. (2007). *Error rates and error types for the Web of Science algorithm for automatic identification of citations*. Presented at the 12th Nordic Workshop on Bibliometrics and Research Policy. 13-14 September 2007, Copenhagen, Denmark.

Lee, Y. W., Pipino, L. L., Funk, J. D., & Wang, R. Y. (2006). *Journey to data quality.* Cambridge, Massachusetts: MIT Press.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information & Management, 40*(2), 133–146. http://dx.doi.org/10.1016/s0378-7206(02)00043-5

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady, 10*, 707–710.

Levin, M., Krawczyk, S., Bethard, S., & Jurafsky, D. (2012). Citation-based bootstrapping for large-scale author disambiguation. *Journal of the American Society for Information Science and Technology, 63*(5), 1030–1047. http://dx.doi.org/10.1002/asi.22621

Levine-Clark, M., & Gil, E. L. (2009). A comparative citation analysis of Web of Science, Scopus, and Google Scholar. *Journal of Business & Finance Librarianship, 14*(1), 32–46. http://dx.doi.org/10.1080/08963560802176348

Levitt, J. M., & Thelwall, M. (2009). Citation levels and collaboration within library and information science. *Journal of the American Society for Information Science and Technology, 60*(3), 434–442. http://dx.doi.org/10.1002/asi.21000

Leydesdorff, L. (2008). Caveats for the use of citation indicators in research and journal evaluations. *Journal of the American Society for Information Science and Technology, 59*(2), 278–287. http://dx.doi.org/10.1002/asi.20743

Li, J., Burnham, J. F., Lemley, T., & Britton, R. M. (2010). Citation analysis: Comparison of Web of Science®, Scopus™, SciFinder®, and Google Scholar. *Journal of Electronic Resources in Medical Libraries, 7*(3), 196–217. http://dx.doi.org/10.1080/15424065.2010.505518

Liang, L., Zhong, Z., & Rousseau, R. (2014). Scientists' referencing (mis)behavior revealed by the dissemination network of referencing errors. *Scientometrics*. http://dx.doi.org/10.1007/s11192-014-1275-x

López-Illescas, C., de Moya-Anegón, F., & Moed, H. F. (2008). Comparing bibliometric country-by-country rankings derived from the Web of Science and Scopus: The effect of poorly cited journals in oncology. *Journal of Information Science, 35*(2), 244–256. http://dx.doi.org/10.1177/0165551508098603

Loshin, D. (2001). *Enterprise knowledge management: The data quality approach.* San Diego: Morgan Kaufmann.

Lundberg, J. (2007). Lifting the crown – citation z-score. *Journal of Informetrics, 1*(2), 145–154. http://dx.doi.org/10.1016/j.joi.2006.09.007

Lynch, B. P., & Smith, K. R. (2001). The changing nature of work in academic libraries. *College & Research Libraries, 62*(5), 407–420. http://dx.doi.org/10.5860/crl.62.5.407

MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science, 40*(5), 342–349. http://dx.doi.org/10.1002/(sici)1097-4571(198909)40:5<342::aid-asi7>3.0.co;2-u

Marsh, E. E., & White, M. D. (2003). A taxonomy of relationships between images and text. *Journal of Documentation, 59*(6), 647–672. http://dx.doi.org/10.1108/00220410310506303

Maydanchik, A. (2007). *Data quality assessment.* Bradley Beach, N.J.: Technics Publications.

Meho, L. I., & Rogers, Y. (2008). Citation counting, citation ranking, and *h*-index of human-computer interaction researchers: A comparison of Scopus and Web of Science. *Journal of the American Society for Information Science and Technology, 59*(11), 1711–1726. http://dx.doi.org/10.1002/asi.20874

Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science vs. Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology, 58*(13), 2105–2125. http://dx.doi.org/10.1002/asi.20677

Meyer, C. A. (2008). Reference accuracy: Best practices for making the links. *Journal of Electronic Publishing, 11*(2). http://dx.doi.org/10.3998/3336451.0011.206

Mingers, J., & Lipitakis, E. A. C. G. (2010). Counting the citations: A comparison of Web of Science and Google Scholar in the field of business and management. *Scientometrics, 85*(2), 613–625. http://dx.doi.org/10.1007/s11192-010-0270-0

Moed, H. F. (1996). Differences in the construction of SCI based bibliometric indicators among various producers: A first over view. *Scientometrics, 35*(2), 177–191. http://dx.doi.org/10.1007/BF02018476

Moed, H. F. (2002). The impact-factors debate: the ISI's uses and limits. *Nature, 415*(6873), 731–732. http://dx.doi.org/10.1038/415731a

Moed, H. F. (2005). *Citation analysis in research evaluation. Information Science and Knowledge Management: Vol. 9.* Dordrecht: Springer. http://dx.doi.org/10.1007/1-4020-3714-7

Moed, H. F., Burger, W. J. M., Frankfort, J. G., & van Raan, A. F. J. (1985). The use of bibliometric data for the measurement of university research performance. *Research Policy, 14*(3), 131–149. http://dx.doi.org/10.1016/0048-7333(85)90012-5

Moed, H. F., & van Leeuwen, T. N. (1995). Improving the accuracy of Institute for Scientific Information's journal impact factors. *Journal of the American Society for Information Science, 46*(6), 461–467. http://dx.doi.org/10.1002/(SICI)1097-4571(199507)46:6<461::AID-ASI5>3.0.CO;2-G

Moed, H. F., & Vriens, M. (1989). Possible inaccuracies occurring in citation analysis. *Journal of Information Science, 15*(2), 95-107. http://dx.doi.org/10.1177/016555158901500205

Narin, F. (1976). *Evaluative Bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity.* Cherry Hill, N. J.: Computer Horizons.

Naranan, S. (1970). Bradford's law of bibliography of science: An interpretation. *Nature, 227*(5258), 631–632. http://dx.doi.org/10.1038/227631a0

Naumann, F. (2002). *Quality-driven query answering for integrated information systems. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/3-540-45921-9

Neuendorf, K. A. (2002). *The content analysis guidebook.* Thousand Oaks, California: Sage Publications.

Neuhaus, C., & Daniel, H.-D. (2008). Data sources for performing citation analysis: An overview. *Journal of Documentation, 64*(2), 193–210. http://dx.doi.org/10.1108/00220410810858010

Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences' literature. *Journal of Informetrics, 1*(2), 161–169. http://dx.doi.org/10.1016/j.joi.2006.12.001

Oermann, M. H., Cummings, S. L., & Wilmes, N. A. (2001). Accuracy of references in four pediatric nursing journals. *Journal of Pediatric Nursing, 16*(4), 263–268. http://dx.doi.org/10.1053/jpdn.2001.25537

Olensky, M. (2012). How is bibliographic data accuracy assessed? In É. Archambault, Y. Gingras, & V. Larivière (Eds.), *Proceedings of the 17th International Conference on Science and Technology Indicators* (pp. 628–639). Montréal, Canada. Retrieved September 24, 2014 from http://2012.sticonference.org/index.php?page=proc

Olensky, M. (2013). Accuracy assessment for bibliographic data. In J. Gorraiz, E. Schiebel, C. Gumpenberger, M. Hörlesberger, & H. F. Moed (Eds.), *Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics*

(pp. 1850–1853). Vienna, Austria. Retrieved September 24, 2014 from http://www.issi2013.org/proceedings.html

On, B.-W., Lee, D., Kang, J., & Mitra, P. (2005). Comparative study of name disambiguation problem using a scalable blocking-based framework. In M. Marlino, T. Sumner, F. M. Shipman III (Eds.), *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2005* (pp. 344–353). Denver, CO, USA. http://dx.doi.org/10.1145/1065385.1065462

O'Neill, E. T., & Vizine-Goetz, D. (1988). Quality control in online databases. *Annual Review of Information Science and Technology (ARIST), 23*, 125–156.

Patton, M. Q. (1999). Enhancing the quality and credibility of qualitative analysis. *Health services research, 34*(5 Pt 2), 1189–1208.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, California: Sage Publications.

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211–218. http://dx.doi.org/10.1145/505248.506010

Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A manifesto, (v.1.0)*. Retrieved August 05, 2014 from http://altmetrics.org/manifesto

Redman, T. C. (1996). *Data quality for the information age. The Artech House computer science library.* Boston, Massachusetts: Artech House.

Reedijk, J. (1998). Sense and nonsense of science citation analyses: Comments on the monopoly position of ISI and citation inaccuracies. Risks of possible misuse and biased citation and impact data. *New Journal of Chemistry, 22*(8), 767–770. http://dx.doi.org/10.1039/a802808g

Research Excellence Framework (2014). *Research Excellence Framework.* Retrieved September 23, 2014 from http://www.ref.ac.uk/

Rittberger, M., & Rittberger, W. (1997). Measuring quality in the production of databases. *Journal of Information Science, 23*(1), 25–37. http://dx.doi.org/10.1177/016555159702300103

Rousseau, R. (1988). Lotka's law and its Leimkuhler representation. *Library science with a slant to documentation and information studies, 25*(3), 150–178.

Rousseau, R. (2002). George Kingsley Zipf: Life, idea, his law and informetrics. *Glottometrics, 3*, 11–18.

Sanderson, M. (2008). Revisiting *h* measured on UK LIS and IR academics. *Journal of the American Society for Information Science and Technology, 59*(7), 1184–1190. http://dx.doi.org/10.1002/asi.20771

Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., & Baldoni, R. (2004). The DaQuinCIS architecture: A platform for exchanging and improving data quality in cooperative information systems. *Information Systems, 29*(7), 551–582. http://dx.doi.org/10.1016/j.is.2003.12.004

Schmidt, M. (2012). Development and evaluation of a match key for linking references to cited articles In É. Archambault, Y. Gingras, & V. Larivière (Eds.), *Proceedings of the 17th International Conference on Science and Technology Indicators* (pp. 707–718). Montréal, Canada. Retrieved September 24, 2014 from http://2012.sticonference.org/index.php?page=proc

Schreier, M. (2012). *Qualitative content analysis in practice.* Los Angeles: Sage Publications.

Simkin, M. V., & Roychowdhury, V. P. (2003). Read before you cite! *Complex Systems, 14*(3), 269–274. Retrieved September 24, 2014 from http://www.complex-systems.com/issues/14-3.html

Slater, M. (1990). *Research methods in library and information studies.* London: Library Association.

Steele, C., Butler, L., & Kingsley, D. (2006). The publishing imperative: The pervasive influence of publication metrics. *Learned Publishing, 19*(4), 277–290. http://dx.doi.org/10.1087/095315106778690751

Strotmann, A., Zhao, D., & Bubela, T. (2009). Author name disambiguation for collaboration network analysis and visualization. *Proceedings of the American Society for Information Science and Technology, 46*(1), 1–20. http://dx.doi.org/10.1002/meet.2009.1450460218

Su, Y., & Jin, Z. (2004). A methodology for information quality assessment in the designing and manufacturing processes of mechanical products. In I. N. Chengalur-Smith, L. Raschid, J. Long, & C. Seko (Eds.), *Proceedings of the Ninth International Conference on Information Quality (ICIQ-04)* (pp. 447–465).

Sweetland, J. H. (1989). Errors in bibliographic citations: A continuing problem. *The library quarterly, 59*(4), 291–304. http://dx.doi.org/10.1086/602160

Synnestvedt, M. B. (2007). *Data preparation for biomedical knowledge domain visualization: A probabilistic record linkage and information fusion approach to citation data*. Doctoral thesis, Drexel University. Retrieved July 11, 2014 from https://idea.library.drexel.edu/islandora/object/idea%3A2532

Tague-Sutcliffe, J. (1992). An introduction to informetrics. *Information processing & management*, *28*(1), 1-3. http://dx.doi.org/10.1016/0306-4573(92)90087-G

Tenopir, C. (1995). Priorities of quality. In R. Basch (Ed.), *Electronic information delivery. Ensuring quality and value* (pp. 119–139). Aldershot, Hampshire, England, Brookfield, Vt., USA: Gower.

Thomson Reuters (2014a). *Web of Science Core Collection.* Retrieved July 11, 2014 from http://wokinfo.com/products_tools/multidisciplinary/webofscience/

Thomson Reuters (2014b). *Web of Science Core Collection Brochure.* Retrieved July 11, 2014 from http://thomsonreuters.com/products/ip-science/04_064/wos-core-collection.pdf

Tunger, D., Haustein, S., Ruppert, L., Luca, G., & Unterhalt, S. (2010). "The Delphic Oracle": An analysis of potential error sources in bibliographic databases. In CWTS (Ed.), *Proceedings of the 11th International Conference on Science and Technology Indicators* (pp. 282–283). Leiden, Netherlands.

Turner, N. B., & Beck, S. E. (2002, June 16). *Search and rescue: Repair strategies of remote users searching the online catalog.* Presentation at the American Library Association Annual Conference, Eighth Annual Reference Research Forum. Atlanta, GA, USA. Retrieved August 19, 2014 from http://surface.syr.edu/sul/74

van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics, 62*(1), 133–143. http://dx.doi.org/10.1007/s11192-005-0008-6

Vaughan, L., & Shaw, D. (2008). A new look at evidence of scholarly citation in citation indexes and from web sources. *Scientometrics, 74*(2), 317–330. http://dx.doi.org/10.1007/s11192-008-0220-2

Velden, T. A., Haque, A.-u., & Lagoze, C. (2011). Resolving author name homonymy to improve resolution of structures in co-author networks. In G. Newton, M. Wright & L. Cassel (Eds.), *Proceedings of the 11th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL 2011* (pp. 241–250). Ottawa, Canada. http://dx.doi.org/10.1145/1998076.1998122

Vieira, E. S., & Gomes, J. A. N. F. (2009). A comparison of Scopus and Web of Science for a typical university. *Scientometrics, 81*(2), 587–600. http://dx.doi.org/10.1007/s11192-009-2178-0

Wallin, J. A. (2005). Bibliometric methods: Pitfalls and possibilities. *Basic & Clinical Pharmacology & Toxicology, 97*(5), 261–275. http://dx.doi.org/10.1111/j.1742-7843.2005.pto_139.x

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: An empirical analysis. *Scientometrics, 87*(3), 467–481. http://dx.doi.org/10.1007/s11192-011-0354-5

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86–95. http://dx.doi.org/10.1145/240455.240479

Wang, R. Y. (1998). A product perspective on total data quality management. *Communications of the ACM, 41*(2), 58–65. http://dx.doi.org/10.1145/269012.269022

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33. Retrieved from September 24, 2014, http://www.jstor.org/stable/40398176

Whitley, K. M. (2002). Analysis of SciFinder Scholar and Web of Science citation searches. *Journal of the American Society for Information Science and Technology, 53*(14), 1210–1215. http://dx.doi.org/10.1002/asi.10192

Winkler, W. E. (1995). Matching and record linkage. In B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge, & P. S. Kott (Eds.), *Business Survey Methods* (pp.355–384). New York: Wiley. http://dx.doi.org/10.1002/9781118150504.ch20

Yannakoudakis, E. J., Ayres, F. H., & Huggill, J. A. W. (1990). Matching of citations between non-standardized databases. *Journal of the American Society for Information Science, 41*(8), 599–610.

Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics*. http://dx.doi.org/10.1007/s11192-014-1264-0

Appendices

A Citation matching algorithms of the applied bibliometric research groups

Information acquired from the applied bibliometric research groups about their citation matching algorithms (M. Schmidt for iFQ, personal communication, August 10, 2014; N.J.P. van Eck for CWTS, personal communication, September 8, 2014)[50]:

- What bibliographic fields do you use to match the citations to their cited articles?

iFQ: citation data of the WoS tagged data format: ca [first author]; cw [source title abbreviation]; cv [volume]; cp [first page]; cy [publication year]; rs_doi [doi]; item, author and source data of the WoS tagged data format: name [first author]; j2, j1, ji, j9 [source title abbreviations] and so [source title] if j2 is not available; vl [volume]; bp [first page], py [pubyear]; ar_doi [doi]

CWTS: We try to use as many information as is available in the WoS data. For the citing references this means: Name of first author; Abbreviated source title; Publication year; Volume number; Beginning page number (or article number). And for the citing publications: Author names; Source title; Publication year; Volume number; Page numbers (or article number).

- Do you base your citation analyses on data from WoS only or do you use complementary data sources (if so, which are those or does it depend on the goal of the study, etc.)?

iFQ: We use WoS data for all of our scientific and service studies. We use TR citation data (R9/T9 links) for all service and scientific studies so far because of reproducibility issues

[50] Unfortunately we did not receive the responses of Science-Metrix in time to include them here.

as well as because of our own algorithm has not really been finished and sufficiently evaluated until now. The own iFQ matching algorithm was planned to be used mainly for the goal of estimating the faultiness of the TR data in the context of a broader project dedicated to an error calculus of bibliometric data.

CWTS: No, we don't use complementary data sources.

- Do you use string matching methodologies in your algorithm? If so, which?

iFQ: A CPAN package for the Damerau-Levenshtein distance (http://search.cpan.org/~ugexe/Text-Levenshtein-Damerau-0.41/lib/Text/Levenshtein/Damerau.pm) is used for source and author names, accompanied by thresholds for performance reasons and, in case of author names, for reliability reasons as well. Apart from that, the algorithm works with abbreviations of one or two characters for both types of strings in single match keys.

CWTS: Yes, we use string matching functions in our algorithm. At some point we apply a fuzzy matching based on the source title using Levenstein [sic] distance. Furthermore, if a perfect match could not be made based on the last name of the first author, then we try to match based on the soundex code of the last name.

- Do you use other resources (e.g. list of ISO abbreviations of journal titles) in the matching process?

iFQ: According to our documentation, the WoS ji field contains ISO source abbreviations. No external data are used.

CWTS: No, we don't use other resources.

B List of the 300 cited articles

Internal ID	Reference of cited article
BeSo03_001	Diewald, M. (2003). Kapital oder Kompensation? Erwerbsbiografien von Männern und die sozialen Beziehungen zu Verwandten und Freunden. *Berliner Journal für Soziologie, 13*(2), 213–238. http://dx.doi.org/10.1007/BF03204576
BeSo03_002	Lohr, K. (2003). Subjektivierung von Arbeit. Ausgangspunkt einer Neuorientierung der Industrie- und Arbeitssoziologie? *Berliner Journal für Soziologie, 13*(4), 511–529. http://dx.doi.org/10.1007/BF03204689
BeSo03_003	Delhey, J. (2003). Europäische Integration, Modernisierung und Konvergenz. Zum Einfluss der EU auf die Konvergenz der Mitgliedsländer. *Berliner Journal für Soziologie, 13*(4), 565–584. http://dx.doi.org/10.1007/BF03204692
BeSo03_004	Goldthorpe, J. H. (2003). Globalisierung und soziale Klassen. *Berliner Journal für Soziologie, 13*(3), 301–323. http://dx.doi.org/10.1007/BF03204672
BeSo03_005	Krüger, H. (2003). Berufliche Bildung. Der deutsche Sonderweg und die Geschlechterfrage. *Berliner Journal für Soziologie, 13*(4), 497–510. http://dx.doi.org/10.1007/BF03204688
BeSo03_006	Wendt, C. (2003). Vertrauen in Gesundheitssysteme. *Berliner Journal für Soziologie, 13*(3), 371–393. http://dx.doi.org/10.1007/BF03204675
BeSo03_007	Deutschmann, C. (2003). Industriesoziologie als Wirklichkeitswissenschaft. *Berliner Journal für Soziologie, 13*(4), 477–495. http://dx.doi.org/10.1007/BF03204687
BeSo03_008	Kühl, S. (2003). New Economy, Risikokapital und die Mythen des Internet. *Berliner Journal für Soziologie, 13*(1), 77–96. http://dx.doi.org/10.1007/BF03204084
BeSo03_009	Lang, F. R. (2003). Die Gestaltung und Regulation sozialer Beziehungen im Lebenslauf: Eine entwicklungspsychologische Perspektive. *Berliner Journal für Soziologie, 13*(2), 175–195. http://dx.doi.org/10.1007/BF03204574
BeSo03_010	Merkel, W. (2003). Institutionen und Reformpolitik: Drei Fallstudien zur Vetospieler -Theorie. *Berliner Journal für Soziologie, 13*(2), 255–274. http://dx.doi.org/10.1007/BF03204578
BeSo08_001	Dolata, U. (2008). Das Internet und die Transformation der Musikindustrie. Rekonstruktion und Erklärung eines unkontrollierten Wandels. *Berliner Journal für Soziologie, 18*(3), 344–369. http://dx.doi.org/10.1007/s11609-008-0025-4
BeSo08_002	Rössel, J. (2008). Vom rationalen Akteur zum „systemic dope“. Eine Auseinandersetzung mit der Sozialtheorie von Hartmut Esser. *Berliner Journal für Soziologie, 18*(1), 156–178. http://dx.doi.org/10.1007/s11609-008-0008-5

BeSo08_003	Söhn, J. (2008). Bildungsunterschiede zwischen Migrantengruppen in Deutschland: Schulabschlüsse von Aussiedlern und anderen Migranten der ersten Generation im Vergleich. *Berliner Journal für Soziologie, 18*(3), 401–431. http://dx.doi.org/10.1007/s11609-008-0028-1
BeSo08_004	Henninger, A., Wimbauer, C., & Dombrowski, R. (2008). Geschlechtergleichheit oder „exklusive Emanzipation“? Ungleichheitssoziologische Implikationen der aktuellen familienpolitischen Reformen. *Berliner Journal für Soziologie, 18*(1), 99–128. http://dx.doi.org/10.1007/s11609-008-0006-7
BeSo08_005	Vobruba, G. (2008). Die Entwicklung der Europasoziologie aus der Differenz national/europäisch. *Berliner Journal für Soziologie, 18*(1), 32–51. http://dx.doi.org/10.1007/s11609-008-0003-x
BeSo08_006	Hadjar, A., Haunberger, S., & Schubert, F. (2008). Bildung und subjektives Wohlbefinden im Zeitverlauf, 1984–2002. Eine Mehrebenenanalyse. *Berliner Journal für Soziologie, 18*(3), 370–400. http://dx.doi.org/10.1007/s11609-008-0027-2
BeSo08_007	Jürgens, K. (2008). Reproduktion als Praxis. Zum Vermittlungszusammenhang von Arbeits- und Lebenskraft. *Berliner Journal für Soziologie, 18*(2), 193–220. http://dx.doi.org/10.1007/s11609-008-0014-7
BeSo08_008	Legnaro, A. (2008). Arbeit, Strafe und der Freiraum der Subjekte. *Berliner Journal für Soziologie, 18*(1), 52–72. http://dx.doi.org/10.1007/s11609-008-0004-9
BeSo08_009	Schwinn, T. (2008). Nationale und globale Ungleichheit. *Berliner Journal für Soziologie, 18*(1), 8–31. http://dx.doi.org/10.1007/s11609-008-0002-y
BeSo08_010	Theobald, H. (2008). Care-Politiken, Care-Arbeitsmarkt und Ungleichheit: Schweden, Deutschland und Italien im Vergleich. *Berliner Journal für Soziologie, 18*(2), 257–281. http://dx.doi.org/10.1007/s11609-008-0018-3
BeSo98_001	Hollstein, B., & Bria, G. (1998). Reziprozität in Eltern-Kind-Beziehungen? Theoretische Überlegungen und empirische Evidenz. *Berliner Journal für Soziologie, 8*(1), 7–22.
BeSo98_002	Offe, C. (1998). Der deutsche Wohlfahrtsstaat: Prinzipien, Leistungen, Zukunftsaussichten. *Berliner Journal für Soziologie, 8*(3), 359–380.
BeSo98_003	Stichweh, R. (1998). Zur Theorie der politischen Inklusion. *Berliner Journal für Soziologie, 8*(4), 539–547.
BeSo98_004	Albrow, M. (1998). Europa im globalen Zeitalter. *Berliner Journal für Soziologie, 8*(3), 411–420.
BeSo98_005	Berking, H. (1998). „Global Flows and Local Cultures“. Über die Rekonfiguration sozialer Räume im Globalisierungsprozeß. *Berliner Journal für Soziologie, 8*(3), 381–392.
BeSo98_006	Bodemann, Y. M. (1998). Von Berlin nach Chicago und weiter. Georg Simmel und die Reise seines „Fremden". *Berliner Journal für Soziologie, 8*(1), 125–142.
BeSo98_007	Sassen, S. (1998). Zur Einbettung des Globalisierungsprozesses: Der Nationalstaat vor neuen Aufgaben. *Berliner Journal für Soziologie, 8*(3), 345–357.
BeSo98_008	Somers, M. R. (1998). "Citizenship" zwischen Staat und Markt. Das Konzept der Zivilgesellschaft und das Problem der "dritten Sphäre". *Berliner Journal für Soziologie, 8*(4), 489–505.

BeSo98_009	Western, B., & Beckett, K. (1998). Der Mythos des freien Marktes. Das Strafrecht als Institution des US-amerikanischen Arbeitsmarktes. *Berliner Journal für Soziologie, 8*(2), 159–180.
BeSo98_010	Willems, H. (1998). Elemente einer Theorie der Theatralität "unanständigen" Verhaltens. *Berliner Journal für Soziologie, 8*(2), 201–222.
HAC03_001	Pardasani, R. T., Pardasani, P., Chaturvedi, V., Yadav, S. K., Saxena, A., & Sharma, I. (2003). Theoretical and synthetic approach to novel spiroheterocycles derived from isatin derivatives and L-proline via 1,3-dipolar cycloaddition. *Heteroatom Chemistry, 14*(1), 36–41. http://dx.doi.org/10.1002/hc.10063
HAC03_002	BelBruno, J. J. (2003). Bonding and energetics in small clusters of gallium and arsenic. *Heteroatom Chemistry, 14*(2), 189–196. http://dx.doi.org/10.1002/hc.10127
HAC03_003	Norkus, E., Stalnionienė, I., & Crans, D. C. (2003). Interaction of pyridine- and 4-hydroxypyridine-2,6-dicarboxylic acids with heavy metal ions in aqueous solutions. *Heteroatom Chemistry, 14*(7), 625–632. http://dx.doi.org/10.1002/hc.10203
HAC03_004	Huang, X., & Xu, J. (2003). Stereospecific synthesis of azeto[2,1-d]-[1,5]benzothiazepin/diazepin-1-ones. *Heteroatom Chemistry, 14*(6), 564–569. http://dx.doi.org/10.1002/hc.10196
HAC03_005	Hassan, A. A., Mourad, A.-F. E., El-Shaieb, K. M., Abou-Zied, A. H., & Döpp, D. (2003). Thermolysis of symmetrical dithiobiurea and thioureidoethylthiourea derivatives. *Heteroatom Chemistry, 14*(6), 535–541. http://dx.doi.org/10.1002/hc.10188
HAC03_006	Arslan, M., Aslan, F., & Ozturk, A. I. (2003). Arylation reaction of N-dichlorophosphoryl-P-trichlorophosphazene. *Heteroatom Chemistry, 14*(2), 138–143. http://dx.doi.org/10.1002/hc.10114
HAC03_007	Shi, D.-Q., Sheng, Z.-L., Liu, X.-P., & Wu, H. (2003). Unsaturated cyclic α-hydroxyphosphonates. *Heteroatom Chemistry, 14*(3), 266–268. http://dx.doi.org/10.1002/hc.10139
HAC03_008	Dimukhametov, M. N., Bajandina, E. V., Davydova, E. Y., Litvinov, I. A., Gubaidullin, A. T., Dobrynin, A. B., Zyablikova, T. A. & Alfonsov, V. A. (2003). Stereoselective synthesis of 1,4,2-oxazaphosphorines as precursors of chiral α-aminophosphonic acids by intramolecular heterocyclization of β-aldiminoalkylphosphites. *Heteroatom Chemistry, 14*(1), 56–61. http://dx.doi.org/10.1002/hc.10054
HAC03_009	Basu Baul, T. S., Dutta, S., Masharing, C., Rivarola, E., & Englert, U. (2003). Organotin (IV) complexes of N-[(2Z)-3-hydroxy-1-methyl-2-butenylidene]glycine. *Heteroatom Chemistry, 14*(2), 149–154. http://dx.doi.org/10.1002/hc.10116
HAC03_010	Keglevich, G., Szelke, H., Bálint, A., Imre, T., Ludányi, K., Nagy, Z., Hanusz, M., Simon, K., Harmat, V. & Tőke, L. (2003). Fragmentation-related phosphinylations using 2-aryl-2-phosphabicyclo[2.2.2]oct-5-ene- and -octa-5,7-diene 2-oxides. *Heteroatom Chemistry, 14*(5), 443–451. http://dx.doi.org/10.1002/hc.10176
HAC98_001	Attaby, F. A., Eldin, S. M., & Elneairy, M. A. (1998). Reactions and Characterization of pyridin-6-one-2-thione and 3-diazotized amino-4-hydroxypyrazolo-[3,4-b]pyridin-6-one. *Heteroatom Chemistry, 9*(6), 571–579. http://dx.doi.org/10.1002/(SICI)1098-1071(1998)9:6<571::AID-HC8>3.0.CO;2-7

HAC98_002	Raghunathan, R., Shanmugasundaram, M., Bhanumathi, S., & Padma Malar, E. J. (1998). Synthesis of spiropyrazoline[5.3']4'-chromanones. *Heteroatom Chemistry, 9*(3), 327–332. http://dx.doi.org/10.1002/(SICI)1098-1071(1998)9:3<327::AID-HC9>3.0.CO;2-6
HAC98_003	Novosad, J., Lindeman, S. V., Marek, J., Woollins, J. D., & Husebye, S. (1998). Synthesis and structural characterization of [Te{$(SePPh_2)_2N$}$_2$] and [4-MeOPhTe{$(SPPh_2)_2N$}]$_2$. *Heteroatom Chemistry, 9*(7), 615–621. http://dx.doi.org/10.1002/(SICI)1098-1071(1998)9:7<615::AID-HC4>3.0.CO;2-1
HAC98_004	Hameed, S., Ahmad, R., & Duddeck, H. (1998). Chiral recognition of selenides and iodides by ^{1}H NMR spectroscopy in the presence of a chiral dirhodium complex. *Heteroatom Chemistry, 9*(5), 471–474. http://dx.doi.org/10.1002/(SICI)1098-1071(1998)9:5<471::AID-HC2>3.0.CO;2-9
HAC98_005	Heinicke, J., & Oprea, A. (1998). Higher carbene homologues: Naphtho[2,3-*d*]-1,3,2 λ²-diazagermole, -diazastannole, and attempted reduction of 2,2-dichloronaphtho[2,3-*d*]-1,3,2-diazasilole. *Heteroatom Chemistry, 9*(4), 439–444. http://dx.doi.org/10.1002/(SICI)1098-1071(1998)9:4<439::AID-HC13>3.0.CO;2-S
HAC98_006	Denmark, S. E., Swiss, K. A., Miller, P. C., & Wilson, S. R. (1998). Solution and solid-state structures of lithiated cyclic phosphonates. *Heteroatom Chemistry, 9*(2), 209–218. http://dx.doi.org/10.1002/(SICI)1098-1071(1998)9:2<209::AID-HC17>3.0.CO;2-V
HAC98_007	Tran Huy, N. H., Ricard, L., & Mathey, F. (1998). Reaction of terminal phosphinidene complexes with aldimines: Synthesis of the first 1,2,3-azadiphosphetidine. *Heteroatom Chemistry, 9*(6), 597–600. http://dx.doi.org/10.1002/(SICI)1098-1071(1998)9:6<597::AID-HC12>3.0.CO;2-M
HAC98_008	Touaux, B., Texier-Boullet, F., & Hamelin, J. (1998). Synthesis of oximes, conversion to nitrile oxides and their subsequent 1,3-dipolar cycloaddition reactions under microwave irradiation and solvent-free reaction conditions. *Heteroatom Chemistry, 9*(3), 351–354. http://dx.doi.org/10.1002/(SICI)1098-1071(1998)9:3<351::AID-HC13>3.0.CO;2-Q
HAC98_009	Gupta, N., Jain, C. B., Heinicke, J., Bharatiya, N., Bansal, R. K., & Jones, P. G. (1998). 2-phosphaindolizines. *Heteroatom Chemistry, 9*(3), 333–339. http://dx.doi.org/10.1002/(SICI)1098-1071(1998)9:3<333::AID-HC10>3.0.CO;2-S
HAC98_010	Camacho-Camacho, C., Tlahuext, H., Nöth, H., & Contreras, R. (1998). Two new dibenzobicyclic penta- and hexacoordinated tin compounds. *Heteroatom Chemistry, 9*(3), 321–326. http://dx.doi.org/10.1002/(SICI)1098-1071(1998)9:3<321::AID-HC8>3.0.CO;2-C
HaCl03_001	Altenmüller, E. (2003). Focal dystonia: advances in brain imaging and understanding of fine motor control in musicians. *Hand Clinics, 19*(3), 523–538. http://dx.doi.org/10.1016/S0749-0712(03)00043-X
HaCl03_002	Schuind, F. A., Mouraux, D., Robert, C., Brassinne, E., Rémy, P., Salvia, P., Meyer, A., Moulart, F. & Burny, F. (2003). Functional and outcome evaluation of the hand and wrist. *Hand Clinics, 19*(3), 361–369. http://dx.doi.org/10.1016/S0749-0712(03)00026-X
HaCl03_003	Brandfonbrener, A. G. (2003). Musculoskeletal problems of instrumental musicians. *Hand Clinics, 19*(2), 231–239. http://dx.doi.org/10.1016/S0749-

	0712(02)00100-2
HaCl03_004	Meyer, T. M. (2003). Psychological aspects of mutilating hand injuries. *Hand Clinics, 19*(1), 41–49. http://dx.doi.org/10.1016/S0749-0712(02)00056-2
HaCl03_005	Giessler, G. A., Erdmann, D., & Germann, G. (2003). Soft tissue coverage in devastating hand injuries. *Hand Clinics, 19*(1), 63–71. http://dx.doi.org/10.1016/S0749-0712(02)00128-2
HaCl03_006	Wei, F.-C., Jain, V., & Chen, S. H.-T. (2003). Toe-to-hand transplantation. *Hand Clinics, 19*(1), 165–175. http://dx.doi.org/10.1016/S0749-0712(02)00127-0
HaCl03_007	Rosén, B., & Lundborg, G. (2003). A new model instrument for outcome after nerve repair. *Hand Clinics, 19*(3), 463–470. http://dx.doi.org/10.1016/S0749-0712(03)00003-9
HaCl03_008	Chamagne, P. (2003). Functional dystonia in musicians: rehabilitation. *Hand Clinics, 19*(2), 309–316. http://dx.doi.org/10.1016/S0749-0712(03)00025-8
HaCl03_009	Tubiana, R. (2003). Musician's focal dystonia. *Hand Clinics, 19*(2), 303–308. http://dx.doi.org/10.1016/S0749-0712(02)00099-9
HaCl03_010	Neumeister, M. W., & Brown, R. E. (2003). Mutilating hand injuries: principles and management. *Hand Clinics, 19*(1), 1–15. http://dx.doi.org/10.1016/S0749-0712(02)00141-5
HaCl98_001	Hargens, A. R., & Mubarak, S. J. (1998). Current concepts in the pathophysiology, evaluation, and diagnosis of compartment syndrome. *Hand Clinics, 14*(3), 371–383.
HaCl98_002	Garcia-Elias, M. (1998). Soft-tissue anatomy and relationships about the distal ulna. *Hand Clinics, 14*(2), 165–176.
HaCl98_003	Gonzalez, M. H. (1998). Necrotizing fasciitis and gangrene of the upper extremity. *Hand Clinics, 14*(4), 635–645.
HaCl98_004	Yamaguchi, S., & Viegas, S. F. (1998). Causes of upper extremity compartment syndrome. *Hand Clinics, 14*(3), 365–370.
HaCl98_005	Ouellette, E. A. (1998). Compartment syndromes in obtunded patients. *Hand Clinics, 14*(3), 431–450.
HaCl98_006	Boles, S. D., & Schmidt, C. C. (1998). Pyogenic flexor tenosynovitis. *Hand Clinics, 14*(4), 567–578.
HaCl98_007	Szabo, R. M. (1998). Acute carpal tunnel syndrome. *Hand Clinics, 14*(3), 419–429.
HaCl98_008	Lichtman, D. M., Ganocy, T. K., & Kim, D. C. (1998). The indications for and techniques and outcomes of ablative procedures of the distal ulna. The Darrach resection, hemiresection, matched resection, and Sauvé-Kapandji procedure. *Hand Clinics, 14*(2), 265–277.
HaCl98_009	James, M. A., & Durkin, R. C. (1998). Nonvascularized toe proximal phalanx transfers in the treatment of aphalangia. *Hand Clinics, 14*(1), 1–15.
HaCl98_010	Murray, P. M. (1998). Septic arthritis of the hand and wrist. *Hand Clinics, 14*(4), 579–587.
JCuSt03_001	de Castell, S., & Jenson, J. (2003). Serious play. *Journal of Curriculum Studies, 35*(6), 649–665. http://dx.doi.org/10.1080/0022027032000145552
JCuSt03_002	Spillane, J. P., Diamond, J. B., & Jita, L. (2003). Leading instruction: the distribution of leadership for instruction. *Journal of Curriculum Studies, 35*(5), 533–543. http://dx.doi.org/10.1080/0022027021000041972

JCuSt03_003	Roth, W.-M. (2003). Scientific literacy as an emergent feature of collective human praxis. *Journal of Curriculum Studies, 35*(1), 9–23. http://dx.doi.org/10.1080/00220270210134600
JCuSt03_004	Wraga, W. G., & Hlebowitsh, P. S. (2003). Toward a renaissance in curriculum theory and development in the USA. *Journal of Curriculum Studies, 35*(4), 425–437. http://dx.doi.org/10.1080/00220270305527
JCuSt03_005	Terhart, E. (2003). Constructivism and teaching: a new paradigm in general didactics? *Journal of Curriculum Studies, 35*(1), 25–44. http://dx.doi.org/10.1080/00220270210163653
JCuSt03_006	Benavot, A., & Resh, N. (2003). Educational governance, school autonomy, and curriculum implementation: a comparative study of Arab and Jewish schools in Israel. *Journal of Curriculum Studies, 35*(2), 171–196. http://dx.doi.org/10.1080/0022027022000022856
JCuSt03_007	Estola, E., & Elbaz-Luwisch, F. (2003). Teaching bodies at work. *Journal of Curriculum Studies, 35*(6), 697–719. http://dx.doi.org/10.1080/0022027032000129523
JCuSt03_008	Ross, V. (2003). Walking around the curriculum tree: an analysis of a third/fourth-grade mathematics lesson. *Journal of Curriculum Studies, 35*(5), 567–584. http://dx.doi.org/10.1080/0022027032000083560
JCuSt03_009	Hopmann, S. T. (2003). On the evaluation of curriculum reforms. *Journal of Curriculum Studies, 35*(4), 459–478. http://dx.doi.org/10.1080/00220270305520
JCuSt03_010	Wilkins, J. L. M., Graham, G., Parker, S., Westfall, S., Fraser, R. G., & Tembo, M. (2003). Time in the arts and physical education and school achievement. *Journal of Curriculum Studies, 35*(6), 721–734. http://dx.doi.org/10.1080/0022027032000035113
JCuSt08_001	Jickling, B., & Wals, A. E. J. (2008). Globalization and environmental education: looking beyond sustainable development. *Journal of Curriculum Studies, 40*(1), 1–21. http://dx.doi.org/10.1080/00220270701684667
JCuSt08_002	Hansen, D. T. (2008). Curriculum and the idea of a cosmopolitan inheritance. *Journal of Curriculum Studies, 40*(3), 289–312. http://dx.doi.org/10.1080/00220270802036643
JCuSt08_003	Ballet, K., & Kelchtermans, G. (2008). Workload and willingness to change: disentangling the experience of intensification. *Journal of Curriculum Studies, 40*(1), 47–67. http://dx.doi.org/10.1080/00220270701516463
JCuSt08_004	Osberg, D., & Biesta, G. (2008). The emergent curriculum: navigating a complex course between unguided learning and planned enculturation. *Journal of Curriculum Studies, 40*(3), 313–328. http://dx.doi.org/10.1080/00220270701610746
JCuSt08_005	Leander, K. M., & Osborne, M. D. (2008). Complex positioning: teachers as agents of curricular and pedagogical reform. *Journal of Curriculum Studies, 40*(1), 23–46. http://dx.doi.org/10.1080/00220270601089199
JCuSt08_006	Sanger, M. N. (2008). What we need to prepare teachers for the moral nature of their work. *Journal of Curriculum Studies, 40*(2), 169–185. http://dx.doi.org/10.1080/00220270701670856
JCuSt08_007	Linn, R. L. (2008). Methodological issues in achieving school accountability. *Journal of Curriculum Studies, 40*(6), 699–711. http://dx.doi.org/10.1080/00220270802105729

JCuSt08_008	Nasser, R., & Nasser, I. (2008). Textbooks as a vehicle for segregation and domination: state efforts to shape Palestinian Israelis' identities as citizens. *Journal of Curriculum Studies, 40*(5), 627–650. http://dx.doi.org/10.1080/00220270802072804
JCuSt08_009	Sloan, K. (2008). The expanding educational services sector: neoliberalism and the corporatization of curriculum at the local level in the US. *Journal of Curriculum Studies, 40*(5), 555–578. http://dx.doi.org/10.1080/00220270701784673
JCuSt08_010	van Driel, J. H., Bulte, A. M. W., & Verloop, N. (2008). Using the curriculum emphasis concept to investigate teachers' curricular beliefs in the context of educational reform. *Journal of Curriculum Studies, 40*(1), 107–122. http://dx.doi.org/10.1080/00220270601078259
JCuSt98_001	Lensmire, T. J. (1998). Rewriting student voice. *Journal of Curriculum Studies, 30*(3), 261–291. http://dx.doi.org/10.1080/002202798183611
JCuSt98_002	Atkin, J. M. (1998). The OECD study of innovations in science, mathematics and technology education. *Journal of Curriculum Studies, 30*(6), 647–660. http://dx.doi.org/10.1080/002202798183369
JCuSt98_003	Taylor, A. (1998). Employability skills: From corporate 'wish list' to government policy. *Journal of Curriculum Studies, 30*(2), 143–164. http://dx.doi.org/10.1080/002202798183675
JCuSt98_004	Reay, D. (1998). Setting the agenda: The growing impact of market forces on pupil grouping in British secondary schooling. *Journal of Curriculum Studies, 30*(5), 545–558. http://dx.doi.org/10.1080/002202798183440
JCuSt98_005	Page, R. N. (1998). Moral aspects of curriculum: 'making kids care' about school knowledge. *Journal of Curriculum Studies, 30*(1), 1–26. http://dx.doi.org/10.1080/002202798183738
JCuSt98_006	French, L., & Song, M. J. (1998). Developmentally appropriate teacher-directed approaches: Images from Korean kindergartens. *Journal of Curriculum Studies, 30*(4), 409–430. http://dx.doi.org/10.1080/002202798183558
JCuSt98_007	Boostrom, R. (1998). 'Safe spaces': Reflections on an educational metaphor. *Journal of Curriculum Studies, 30*(4), 397–408. http://dx.doi.org/10.1080/002202798183549
JCuSt98_008	Upitis, R. (1998). From hackers to luddites, game players to game creators: Profiles of adolescent students using technology. *Journal of Curriculum Studies, 30*(3), 293–318. http://dx.doi.org/10.1080/002202798183620
JCuSt98_009	Segal, S. (1998). The role of contingency and tension in the relationship between theory and practice in the classroom. *Journal of Curriculum Studies, 30*(2), 199–206. http://dx.doi.org/10.1080/002202798183701
JCuSt98_010	Riquarts, K., & Hansen, K.-H. (1998). Collaboration among teachers, researchers and in-service trainers to develop an integrated science curriculum. *Journal of Curriculum Studies, 30*(6), 661–676. http://dx.doi.org/10.1080/002202798183378
JTM03_001	Schlagenhauf, P., Steffen, R., & Loutan, L. (2003). Migrants as a major risk group for imported malaria in European countries. *Journal of Travel Medicine, 10*(2), 106–107. http://dx.doi.org/10.2310/7060.2003.31764
JTM03_002	van Herck, K., Zuckerman, J., Castelli, F., van Damme, P., Walker, E., Steffen, R. & for the European Travel Health Advisory Board (2003). Travelers' knowledge, attitudes, and practices on prevention of infectious diseases: Results from a pilot study. *Journal of Travel Medicine, 10*(2), 75–78. http://dx.doi.org/10.2310/7060.2003.31638

JTM03_003	Grobusch, M. P., Mühlberger, N., Jelinek, T., Bisoffi, Z., Corachán, M., Harms, G., Matteelli, A., Fry, G., Hatz, C., Gjørup, I., Schmid, M. L., Knobloch, J., Puente, S., Bronner, U., Kapaun, A., Clerinx, J., Nielsen, L. N., Fleischer, K., Beran, J., da Cunha, S., Schulze, M., Myrvang, B., & Hellgren, U. (2003). Imported schistosomiasis in Europe: Sentinel surveillance data from TropNetEurop. *Journal of Travel Medicine, 10*(3), 164–169. http://dx.doi.org/10.2310/7060.2003.35759
JTM03_004	Cabada, M. M., Montoya, M., Echevarria, J. I., Verdonck, K., Seas, C., & Gotuzzo, E. (2003). Sexual behavior in travelers visiting Cuzco. *Journal of Travel Medicine, 10*(4), 214–218. http://dx.doi.org/10.2310/7060.2003.40508
JTM03_005	Weber, R., Schlagenhauf, P., Amsler, L., & Steffen, R. (2003). Knowledge, attitudes and practices of business travelers regarding malaria risk and prevention. *Journal of Travel Medicine, 10*(4), 219–224. http://dx.doi.org/10.2310/7060.2003.40574
JTM03_006	Wilder-Smith, A., Goh, K. T., & Paton, N. I. (2003). Experience of severe acute respiratory syndrome in Singapore: Importation of cases, and defense strategies at the airport. *Journal of Travel Medicine, 10*(5), 259–262. http://dx.doi.org/10.2310/7060.2003.2676
JTM03_007	Nigro, L., Larocca, L., Massarelli, L., Patamia, I., Minniti, S., Palermo, F., & Cacopardo, B. (2003). A placebo-controlled treatment trial of Blastocystis hominis infection with metronidazole. *Journal of Travel Medicine, 10*(2), 128–130. http://dx.doi.org/10.2310/7060.2003.31714
JTM03_008	Duval, B., de Serre, G., Shadmani, R., Boulianne, N., Pohani, G., Naus, M., Rochette, L., Fradet, M. D., Kain, K. C. & Ward, B. J. (2003). A population-based comparison between travelers who consulted travel clinics and those who did not. *Journal of Travel Medicine, 10*(1), 4–10. http://dx.doi.org/10.2310/7060.2003.30659
JTM03_009	Salomon, J., Flament Saillour, M., Truchis, P. de, Bougnoux, M. E., Dromer, F., Dupont, B., Saint-Hardouin, G. & Peronne, C. (2003). An outbreak of acute pulmonary histoplasmosis in members of a trekking trip in Martinique, French West Indies. *Journal of Travel Medicine, 10*(2), 87–93. http://dx.doi.org/10.2310/7060.2003.31755
JTM03_010	Thompson, D. T., Ashley, D. V., Dockery-Brown, C. A., Binns, A., Jolly, C. M., & Jolly, P. E. (2003). Incidence of health crises in tourists visiting Jamaica, West Indies, 1998 to 2000. *Journal of Travel Medicine, 10*(2), 79–86. http://dx.doi.org/10.2310/7060.2003.31628
JTM98_001	Phillips-Howard, P. A., Steffen, R., Kerr, L., Vanhauwere, B., Schildknecht, J., Fuchs, E., & Edwards, R. (1998). Safety of mefloquine and other antimalarial agents in the first trimester of pregnancy. *Journal of Travel Medicine, 5*(3), 121–126. http://dx.doi.org/10.1111/j.1708-8305.1998.tb00484.x
JTM98_002	Schoepke, A., Steffen, R., & Gratz, N. (1998). Effectiveness of personal protection measures against mosquito bites for malaria prophylaxis in travelers. *Journal of Travel Medicine, 5*(4), 188–192. http://dx.doi.org/10.1111/j.1708-8305.1998.tb00505.x
JTM98_003	van Hoecke, C., Lebacq, E., Beran, J., Prymula, R., & Collard, F. (1998). Concomitant vaccination against hepatitis A and typhoid fever. *Journal of Travel Medicine, 5*(3), 116–120. http://dx.doi.org/10.1111/j.1708-8305.1998.tb00483.x

JTM98_004	Kemmerer, T. P., Cetron, M., Harper, L., & Kozarsky, P. E. (1998). Health problems of corporate travelers: Risk factors and management. *Journal of Travel Medicine, 5*(4), 184–187. http://dx.doi.org/10.1111/j.1708-8305.1998.tb00504.x
JTM98_005	Gambel, J. M., Brundage, J. F., Kuschner, R. A., & Kelley, P. W. (1998). Deployed US Army soldiers' knowledge and use of personal protection measures to prevent arthropod-related casualties. *Journal of Travel Medicine, 5*(4), 217–220. http://dx.doi.org/10.1111/j.1708-8305.1998.tb00511.x
JTM98_006	McIntosh, I. B., Swanson, V., Power, K. G., Raeside, F., & Dempster, C. (1998). Anxiety and health problems related to air travel. *Journal of Travel Medicine, 5*(4), 198–204. http://dx.doi.org/10.1111/j.1708-8305.1998.tb00507.x
JTM98_007	Durrheim, D. N., Braack, L. E. O., Waner, S., & Gammon, S. (1998). Risk of malaria in visitors to the Kruger National Park, South Africa. *Journal of Travel Medicine, 5*(4), 173–177. http://dx.doi.org/10.1111/j.1708-8305.1998.tb00502.x
JTM98_008	Elawad, B. B., & Ong, E. L. (1998). Retrospective study of malaria cases treated in Newcastle General Hospital between 1990 and 1996. *Journal of Travel Medicine, 5*(4), 193–197. http://dx.doi.org/10.1111/j.1708-8305.1998.tb00506.x
JTM98_009	Gehring, T. M., Widmer, J., Kleiber, D., & Steffen, R. (1998). Are preventive HIV interventions at airports effective? *Journal of Travel Medicine, 5*(4), 205–209. http://dx.doi.org/10.1111/j.1708-8305.1998.tb00508.x
JTM98_010	Weber, G., Borer, A., Zirkin, H. J., Riesenberg, K., & Alkan, M. (1998). Schistosomiasis presenting as acute appendicitis in a traveler. *Journal of Travel Medicine, 5*(3), 147–148. http://dx.doi.org/10.1111/j.1708-8305.1998.tb00489.x
PoTh03_001	Pagden, A. (2003). Human rights, natural rights, and Europe's imperial legacy. *Political Theory, 31*(2), 171–199. http://dx.doi.org/10.1177/0090591702251008
PoTh03_002	Scott, D. (2003). Culture in political theory. *Political Theory, 31*(1), 92–115. http://dx.doi.org/10.1177/0090591702239440
PoTh03_003	McCormick, J. P. (2003). Machiavelli against republicanism. On the Cambridge School's "Guicciardinian moments". *Political Theory, 31*(5), 615–643. http://dx.doi.org/10.1177/0090591703252159
PoTh03_004	Welch, C. B. (2003). Colonial violence and the rhetoric of evasion. Tocqueville on Algeria. *Political Theory, 31*(2), 235–264. http://dx.doi.org/10.1177/0090591702251011
PoTh03_005	Dallmayr, F. (2003). Cosmopolitanism: Moral and political. *Political Theory, 31*(3), 421–442. http://dx.doi.org/10.1177/0090591703251909
PoTh03_006	Abdel-Nour, F. (2003). National responsibility. *Political Theory, 31*(5), 693–719. http://dx.doi.org/10.1177/0090591703252156
PoTh03_007	Bohman, J. (2003). Deliberative toleration. *Political Theory, 31*(6), 757–779. http://dx.doi.org/10.1177/0090591703252379
PoTh03_008	Näsström, S. (2003). What globalization overshadows. *Political Theory, 31*(6), 808–834. http://dx.doi.org/10.1177/0090591703252158
PoTh03_009	Deveaux, M. (2003). A deliberative approach to conflicts of culture. *Political Theory, 31*(6), 780–807. http://dx.doi.org/10.1177/0090591703256685

PoTh03_010	Bader, V. (2003). Religious diversity and democratic institutional pluralism. *Political Theory, 31*(2), 265–294. http://dx.doi.org/10.1177/0090591702251012
PoTh08_001	Abizadeh, A. (2008). Democratic theory and border coercion. No right to unilaterally control your own borders. *Political Theory, 36*(1), 37–65. http://dx.doi.org/10.1177/0090591707310090
PoTh08_002	Markell, P. (2008). The insufficiency of non-domination. *Political Theory, 36*(1), 9–36. http://dx.doi.org/10.1177/0090591707310090
PoTh08_003	Farr, J. (2008). Locke, natural law, and new world slavery. *Political Theory, 36*(4), 495–522. http://dx.doi.org/10.1177/0090591708317899
PoTh08_004	Muthu, S. (2008). Adam Smith's critique of international trading companies: Theorizing "Globalization" in the age of enlightenment. *Political Theory, 36*(2), 185–212. http://dx.doi.org/10.1177/0090591707312430
PoTh08_005	Wendt, A., & Duvall, R. (2008). Sovereignty and the UFO. *Political Theory, 36*(4), 607–633. http://dx.doi.org/10.1177/0090591708317902
PoTh08_006	Luxon, N. (2008). Ethics and subjectivity. Practices of self-governance in the late lectures of Michel Foucault. *Political Theory, 36*(3), 377–402. http://dx.doi.org/10.1177/0090591708315143
PoTh08_007	Nelson, E. (2008). From primary goods to capabilities - Distributive justice and the problem of neutrality. *Political Theory, 36*(1), 93–122. http://dx.doi.org/10.1177/0090591707310088
PoTh08_008	Cohen, J. L. (2008). Rethinking human rights, democracy, and sovereignty in the age of globalization. *Political Theory, 36*(4), 578–606. http://dx.doi.org/10.1177/0090591708317901
PoTh08_009	de Roover, J., & Balagangadhara, S. N. (2008). John Locke, Christian liberty, and the predicament of liberal toleration. *Political Theory, 36*(4), 523–549. http://dx.doi.org/10.1177/0090591708317969
PoTh08_010	Leeb, C. (2008). Toward a theoretical outline of the subject. The centrality of Adorno and Lacan for feminist political theorizing. *Political Theory, 36*(3), 351–376. http://dx.doi.org/10.1177/0090591708315142
PoTh98_001	Honneth, A. (1998). Democracy as reflexive cooperation. John Dewey and the theory of democracy today. *Political Theory, 26*(6), 763–783. http://dx.doi.org/10.1177/0090591798026006001
PoTh98_002	Macedo, S. (1998). Transformative constitutionalism and the case of religion: Defending the moderate hegemony of liberalism. *Political Theory, 26*(1), 56–80. http://dx.doi.org/10.1177/0090591798026001004
PoTh98_003	Zerilli, L. M. G. (1998). Doing without knowing. Feminism's politics of the ordinary. *Political Theory, 26*(4), 435–458. http://dx.doi.org/10.1177/0090591798026004001
PoTh98_004	Pasquino, P. (1998). Locke on king's prerogative. *Political Theory, 26*(2), 198–208. http://dx.doi.org/10.1177/0090591798026002003
PoTh98_005	Kukathas, C. (1998). Liberalism and multiculturalism. The politics of indifference. *Political Theory, 26*(5), 686–699. http://dx.doi.org/10.1177/0090591798026005003
PoTh98_006	Digeser, P. (1998). Forgiveness and politics: Dirty hands and imperfect procedures. *Political Theory, 26*(5), 700–724. http://dx.doi.org/10.1177/0090591798026005004
PoTh98_007	Schmidt, J. (1998). Cabbage heads and gulps of water: Hegel on the terror. *Political Theory, 26*(1), 4–32. http://dx.doi.org/10.1177/0090591798026001002

PoTh98_008	Urbinati, N. (1998). From the periphery of modernity. Antonio Gramsci's theory of subordination and hegemony. *Political Theory, 26*(3), 370–391. http://dx.doi.org/10.1177/0090591798026003005
PoTh98_009	Wokler, R. (1998). Contextualizing Hegel's phenomenology of the French Revolution and the terror. *Political Theory, 26*(1), 33–55. http://dx.doi.org/10.1177/0090591798026001003
PoTh98_010	Critchley, S. (1998). Metaphysics in the dark: A response to Richard Rorty and Ernesto Laclau. *Political Theory, 26*(6), 803–817. http://dx.doi.org/10.1177/0090591798026006003
PoVi03_001	Treib, O. (2003). Die Umsetzung von EU-Richtlinien im Zeichen der Parteipolitik: Eine akteurszentrierte Antwort auf die Misfit-These. *Politische Vierteljahresschrift, 44*(4), 506–528. http://dx.doi.org/10.1007/s11615-003-0096-y
PoVi03_002	Freitag, M., Vatter, A., & Müller, C. (2003). Bremse oder Gaspedal? Eine empirische Untersuchung zur Wirkung der direkten Demokratie auf den Steuerstaat. *Politische Vierteljahresschrift, 44*(3), 348–369. http://dx.doi.org/10.1007/s11615-003-0068-2
PoVi03_003	Behnke, J. (2003). Ein integrales Modell der Ursachen von Überhangmandaten. *Politische Vierteljahresschrift, 44*(1), 41–65. http://dx.doi.org/10.1007/s11615-003-0005-4
PoVi03_004	Westle, B. (2003). Europäische Identifikation im Spannungsfeld regionaler und nationaler Identitäten. Theoretische Überlegungen und empirische Befunde. *Politische Vierteljahresschrift, 44*(4), 453–482. http://dx.doi.org/10.1007/s11615-003-0094-0
PoVi03_005	Plümper, T. (2003). Publikationstätigkeit und Rezeptionserfolg der deutschen Politikwissenschaft in internationalen Fachzeitschriften, 1990–2002. Eine bibliometrische Analyse der Veröffentlichungsleistung deutscher politikwissenschaftlicher Fachbereiche und Institute. *Politische Vierteljahresschrift, 44*(4), 529–544. http://dx.doi.org/10.1007/s11615-003-0097-x
PoVi03_006	Becker, R., & Mays, A. (2003). Soziale Herkunft, politische Sozialisation und Wählen im Lebensverlauf. *Politische Vierteljahresschrift, 44*(1), 19–40. http://dx.doi.org/10.1007/s11615-003-0004-5
PoVi03_007	Borchert, J., & Stolz, K. (2003). Die Bekämpfung der Unsicherheit: Politikerkarrieren und Karrierepolitik in der Bundesrepublik Deutschland. *Politische Vierteljahresschrift, 44*(2), 148–173. http://dx.doi.org/10.1007/s11615-003-0036-x
PoVi03_008	Bussmann, M., Scheuthle, H., & Schneider, G. (2003). Die „Friedensdividende" der Globalisierung: Außenwirtschaftliche Öffnung und innenpolitische Stabilität in den Entwicklungsländern. *Politische Vierteljahresschrift, 44*(3), 302–324. http://dx.doi.org/10.1007/s11615-003-0066-4
PoVi03_009	Hardmeier, S., & Roth, H. (2003). Die Erforschung der Wirkung politischer Meinungsumfragen: Lehren vom "Sonderfall" Schweiz. *Politische Vierteljahresschrift, 44*(2), 174–195. http://dx.doi.org/10.1007/s11615-003-0037-9
PoVi03_010	Schaltegger, C. A., & Feld, L. P. (2003). Die Zentralisierung der Staatstätigkeit in einer Referendumsdemokratie: Evidenz aus der Schweiz. *Politische Vierteljahresschrift, 44*(3), 370–394. http://dx.doi.org/10.1007/s11615-003-0069-1

PoVi08_001	Bräuninger, T., & Debus, M. (2008). Der Einfluss von Koalitionsaussagen, programmatischen Standpunkten und der Bundespolitik auf die Regierungsbildung in den deutschen Ländern. *Politische Vierteljahresschrift, 49*(2), 309–338. http://dx.doi.org/10.1007/s11615-008-0101-6
PoVi08_002	König, T., & Mäder, L. (2008). Das Regieren jenseits des Nationalstaates und der Mythos einer 80-Prozent-Europäisierung in Deutschland. *Politische Vierteljahresschrift, 49*(3), 438–463. http://dx.doi.org/10.1007/s11615-008-0106-1
PoVi08_003	Thaa, W. (2008). Kritik und Neubewertung politischer Repräsentation: vom Hindernis zur Möglichkeitsbedingung politischer Freiheit. *Politische Vierteljahresschrift, 49*(4), 618–640. http://dx.doi.org/10.1007/s11615-008-0116-z
PoVi08_004	Linhart, E., Pappi, F. U., & Schmitt, R. (2008). Die proportionale Ministerienaufteilung in deutschen Koalitionsregierungen: Akzeptierte Norm oder das Ausnutzen strategischer Vorteile? *Politische Vierteljahresschrift, 49*(1), 46–67. http://dx.doi.org/10.1007/s11615-008-0087-0
PoVi08_005	Behnke, J. (2008). Strategisches Wählen bei der Nachwahl in Dresden zur Bundestagswahl 2005. *Politische Vierteljahresschrift, 49*(4), 695–720. http://dx.doi.org/10.1007/s11615-008-0119-9
PoVi08_006	Weihe, A. C., Pritzlaff, T., Nullmeier, F., Felgenhauer, T., & Baumgarten, B. (2008). Wie wird in politischen Gremien entschieden? Konzeptionelle und methodische Grundlagen der Gremienanalyse. *Politische Vierteljahresschrift, 49*(2), 339–359. http://dx.doi.org/10.1007/s11615-008-0102-5
PoVi08_007	Sattler, T., & Walter, S. (2008). Wirtschaftspolitischer Handlungsspielraum im Zeitalter der Globalisierung. Eine empirische Untersuchung am Beispiel von Währungskrisen. *Politische Vierteljahresschrift, 49*(3), 464–490. http://dx.doi.org/10.1007/s11615-008-0107-0
PoVi08_008	Bühlmann, M., Merkel, W., Müller, L., & Weßels, B. (2008). Wie lässt sich Demokratie am besten messen? Zum Forumsbeitrag von Thomas Müller und Susanne Pickel. *Politische Vierteljahresschrift, 49*(1), 114–122. http://dx.doi.org/10.1007/s11615-008-0089-y
PoVi08_009	Schoen, H. (2008). Die Deutschen und die Türkeifrage: eine Analyse der Einstellungen zum Antrag der Türkei auf Mitgliedschaft in der Europäischen Union. *Politische Vierteljahresschrift, 49*(1), 68–91. http://dx.doi.org/10.1007/s11615-008-0090-5
PoVi08_010	Bellucci, P. (2008). Why Berlusconi's Landslide Return? A Comment on the 2008 Italian General Election. *Politische Vierteljahresschrift, 49*(4), 605–617. http://dx.doi.org/10.1007/s11615-008-0115-0
PoVi98_001	Benz, A. (1998). Politikverflechtung ohne Politikverflechtungsfalle - Koordination und Strukturdynamik im europäischen Mehrebenensystem. *Politische Vierteljahresschrift, 39*(3), 558–589.
PoVi98_002	Patzelt, W. J. (1998). Ein latenter Verfassungskonflikt? Die Deutschen und ihr parlamentarisches Regierungssystem. *Politische Vierteljahresschrift, 39*(4), 725–757.
PoVi98_003	Abromeit, H. (1998). Ein Vorschlag zur Demokratisierung des europäischen Entscheidungssystems. *Politische Vierteljahresschrift, 39*(1), 80–90.
PoVi98_004	Braun, D. (1998). Der Einfluß von Ideen und Überzeugungssystemen auf die politische Problemlösung. *Politische Vierteljahresschrift, 39*(4), 797–

	818.
PoVi98_005	Rippl, S., Boehnke, K., Hefler, G., & Hagan, J. (1998). Sind Männer eher rechtsextrem und wenn ja, warum? Individualistische Werthaltungen und rechtsextreme Einstellungen. *Politische Vierteljahresschrift, 39*(4), 758–774.
PoVi98_006	Reißig, R. (1998). Transformationsforschung: Gewinne, Desiderate und Perspektiven. *Politische Vierteljahresschrift, 39*(2), 301–328.
PoVi98_007	Schlichte, K. (1998). Struktur und Prozeß: Zur Erklärung bewaffneter Konflikte im nachkolonialen Afrika südlich der Sahara. *Politische Vierteljahresschrift, 39*(2), 261–281.
PoVi98_008	Genschel, P. (1998). Markt und Staat in Europa. *Politische Vierteljahresschrift, 39*(1), 55–79.
PoVi98_009	Rössel, J. (1998). Der Effekt der Organisation. Mobilisierung und Erfolg von Challenging Groups am Beispiel amerikanischer Kohlenbergarbeiterstreiks 1881 - 1894. *Politische Vierteljahresschrift, 39*(1), 28–54.
PoVi98_010	Manow, P., & Plümper, T. (1998). Die Erkenntnisgrenzen der Diskursanalyse. Ein Kommentar zu Elmar Rieger und Stephan Leibfried. *Politische Vierteljahresschrift, 39*(3), 590–601.
SoIn03_001	Sasson-Levy, O. (2003). Feminism and military gender practices: Israeli women soldiers in "masculine" roles. *Sociological Inquiry, 73*(3), 440–465. http://dx.doi.org/10.1111/1475-682X.00064
SoIn03_002	Aguilera, M. B. (2003). The impact of the worker: How social capital and human capital influence the job tenure of formerly undocumented Mexican immigrants. *Sociological Inquiry, 73*(1), 52–83. http://dx.doi.org/10.1111/1475-682X.00041
SoIn03_003	Hannon, L. (2003). Poverty, delinquency, and educational attainment: Cumulative disadvantage or disadvantage saturation? *Sociological Inquiry, 73*(4), 575–594. http://dx.doi.org/10.1111/1475-682X.00072
SoIn03_004	Knudsen, H. K., Johnson, J. A., Martin, J. K., & Roman, P. M. (2003). Downsizing survival: The experience of work and organizational commitment. *Sociological Inquiry, 73*(2), 265–283. http://dx.doi.org/10.1111/1475-682X.00056
SoIn03_005	Kramer, L. A., & Berg, E. C. (2003). A survival analysis of timing of entry into prostitution: The differential impact of race, educational level, and childhood/adolescent risk factors. *Sociological Inquiry, 73*(4), 511–528. http://dx.doi.org/10.1111/1475-682X.00069
SoIn03_006	Harris, M. A. (2003). Religiosity and perceived future ascetic deviance and delinquency among Mormon adolescents: Testing the "this-worldly" supernatural sanctions thesis. *Sociological Inquiry, 73*(1), 28–51. http://dx.doi.org/10.1111/1475-682X.00040
SoIn03_007	Futrell, R. (2003). Framing processes, cognitive liberation, and NIMBY protest in the U.S. chemical-weapons disposal conflict. *Sociological Inquiry, 73*(3), 359–386. http://dx.doi.org/10.1111/1475-682X.00061
SoIn03_008	Featherstone, R., & Deflem, M. (2003). Anomie and strain: Context and consequences of Merton's two theories. *Sociological Inquiry, 73*(4), 471–489. http://dx.doi.org/10.1111/1475-682X.00067
SoIn03_009	Lomsky-Feder, E., & Rapoport, T. (2003). Juggling models of masculinity: Russian-Jewish immigrants in the Israeli army. *Sociological Inquiry, 73*(1), 114–137. http://dx.doi.org/10.1111/1475-682X.00043

SoIn03_010	Oh, J.-H. (2003). Assessing the social bonds of elderly neighbors: The roles of length of residence, crime victimization, and perceived disorder. *Sociological Inquiry, 73*(4), 490–510. http://dx.doi.org/10.1111/1475-682X.00068
SoIn08_001	Freudenburg, W. R., Gramling, R., & Davidson, D. J. (2008). Scientific Certainty Argumentation Methods (SCAMs): Science and the politics of doubt. *Sociological Inquiry, 78*(1), 2–38. http://dx.doi.org/10.1111/j.1475-682X.2008.00219.x
SoIn08_002	Fulkerson, G. M., & Thompson, G. H. (2008). The Evolution of a Contested Concept: A Meta-Analysis of Social Capital Definitions and Trends (1988-2006). *Sociological Inquiry, 78*(4), 536–557. http://dx.doi.org/10.1111/j.1475-682X.2008.00260.x
SoIn08_003	Marshall, B. K., & Picou, J. S. (2008). Postnormal science, precautionary principle, and worst cases: The challenge of twenty-first century catastrophes. *Sociological Inquiry, 78*(2), 230–247. http://dx.doi.org/10.1111/j.1475-682X.2008.00236.x
SoIn08_004	Huisman, K. (2008). "Does This Mean You're Not Going to Come Visit Me Anymore?": An inquiry into an ethics of reciprocity and positionality in feminist ethnographic research. *Sociological Inquiry, 78*(3), 372–396. http://dx.doi.org/10.1111/j.1475-682X.2008.00244.x
SoIn08_005	Tierney, K. (2008). Hurricane in New Orleans? Who knew? Anticipating Katrina and its devastation. *Sociological Inquiry, 78*(2), 179–183. http://dx.doi.org/10.1111/j.1475-682X.2008.00233.x
SoIn08_006	Choi, K. H., Sakamoto, A., & Powers, D. (2008). Who is Hispanic? Hispanic identity among African Americans, Asian Americans, Others, and whites. *Sociological Inquiry, 78*(3), 335–371. http://dx.doi.org/10.1111/j.1475-682X.2008.00243.x
SoIn08_007	Senier, L. (2008). "It's Your Most Precious Thing": Worst-case thinking, trust, and parental decision making about vaccinations. *Sociological Inquiry, 78*(2), 207–229. http://dx.doi.org/10.1111/j.1475-682X.2008.00235.x
SoIn08_008	Dahlin, E. C., & Hironaka, A. (2008). Citizenship beyond borders: A cross-national study of dual citizenship. *Sociological Inquiry, 78*(1), 54–73. http://dx.doi.org/10.1111/j.1475-682X.2008.00221.x
SoIn08_009	Shriver, T. E., Cable, S., & Kennedy, D. (2008). Mining for Conflict and Staking Claims: Contested Illness at the Tar Creek Superfund Site. *Sociological Inquiry, 78*(4), 558–579. http://dx.doi.org/10.1111/j.1475-682X.2008.00258.x
SoIn08_010	Baird, J., Adelman, R. M., Reid, L. W., & Jaret, C. (2008). Immigrant settlement patterns: The role of metropolitan characteristics. *Sociological Inquiry, 78*(3), 310–334. http://dx.doi.org/10.1111/j.1475-682X.2008.00242.x
SoIn98_001	Calhoun, C. (1998). Community without propinquity revisited: Communications technology and the transformation of the urban public sphere. *Sociological Inquiry, 68*(3), 373–397. http://dx.doi.org/10.1111/j.1475-682X.1998.tb00474.x
SoIn98_002	Sprecher, S., & Regan, P. C. (1998). Passionate and companionate love in courting and young married couples. *Sociological Inquiry, 68*(2), 163–185. http://dx.doi.org/10.1111/j.1475-682X.1998.tb00459.x
SoIn98_003	Hunter, M. L. (1998). Colorstruck: Skin color stratification in the lives of African American women. *Sociological Inquiry, 68*(4), 517–535. http://dx.doi.org/10.1111/j.1475-682X.1998.tb00483.x

SoIn98_004	Shriver, T. E., White, D. A., & Kebede, A. (1998). Power, politics, and the framing of environmental illness. *Sociological Inquiry, 68*(4), 458–475. http://dx.doi.org/10.1111/j.1475-682X.1998.tb00480.x
SoIn98_005	Horton, H. D., & Thomas, M. E. (1998). Race, class, and family structure: Differences in housing values for black and white homeowners. *Sociological Inquiry, 68*(1), 114–136. http://dx.doi.org/10.1111/j.1475-682X.1998.tb00456.x
SoIn98_006	Williams, J. (1998). Knowledge, consequences, and experience: The social construction of environmental problems. *Sociological Inquiry, 68*(4), 476–497. http://dx.doi.org/10.1111/j.1475-682X.1998.tb00481.x
SoIn98_007	Berbrier, M. (1998). White supremacists and the (pan-)ethnic imperative: On "European-Americans" and "White Student Unions". *Sociological Inquiry, 68*(4), 498–516. http://dx.doi.org/10.1111/j.1475-682X.1998.tb00482.x
SoIn98_008	Cerulo, K. A., & Ruane, J. M. (1998). Coming together: New taxonomies for the analysis of social relations. *Sociological Inquiry, 68*(3), 398–425. http://dx.doi.org/10.1111/j.1475-682X.1998.tb00475.x
SoIn98_009	Wilder, E. I., & Walters, W. H. (1998). Ethnic and religious components of the Jewish income advantage, 1969 and 1989. *Sociological Inquiry, 68*(3), 426–436. http://dx.doi.org/10.1111/j.1475-682X.1998.tb00476.x
SoIn98_010	Reid, L. W., Roberts, J. T., & Hilliard, H. M. (1998). Fear of crime and collective action: An analysis of coping strategies. *Sociological Inquiry, 68*(3), 312–328. http://dx.doi.org/10.1111/j.1475-682X.1998.tb00470.x
WCZ03_001	Wasserscheid, P. (2003). Ionische Flüssigkeiten: Innovative Lösungsmittel für die Zweiphasenkatalyse. *Chemie in unserer Zeit, 37*(1), 52–63. http://dx.doi.org/10.1002/ciuz.200390006
WCZ03_002	Schieberle, P., & Hofmann, T. (2003). Die molekulare Welt des Lebensmittelgenusses: Auf den Geschmack gekommen. *Chemie in unserer Zeit, 37*(6), 388–401. http://dx.doi.org/10.1002/ciuz.200300305
WCZ03_003	Schmatloch, S., & Schubert, U. S. (2003). Vom einfachen Komplex zum komplexen Gitter: Metallo-supramolekulare Chemie. *Chemie in unserer Zeit, 37*(3), 180–187. http://dx.doi.org/10.1002/ciuz.200300247
WCZ03_004	Beckmann, M., & Haack, K.-J. (2003). Insektizide für die Landwirtschaft: Chemische Schädlingsbekämpfung. *Chemie in unserer Zeit, 37*(2), 88–97. http://dx.doi.org/10.1002/ciuz.200300268
WCZ03_005	Leitner, W. (2003). Chemische Synthese in überkritischem Kohlendioxid: Die „bessere Lösung"?. *Chemie in unserer Zeit, 37*(1), 32–38. http://dx.doi.org/10.1002/ciuz.200390002
WCZ03_006	Henningsen, M. (2003). Moderne Fungizide: Pilzbekämpfung in der Landwirtschaft. *Chemie in unserer Zeit, 37*(2), 98–111. http://dx.doi.org/10.1002/ciuz.200300283
WCZ03_007	Kreisel, G., Wolf, C., Weigand, W., & Dörr, M. (2003). Wie entstand das Leben auf der Erde? Ammoniak aus Stickstoff unter präbiotischen Bedingungen. *Chemie in unserer Zeit, 37*(5), 306–313. http://dx.doi.org/10.1002/ciuz.200300266
WCZ03_008	Sur, R., Hajimiragha, H., Begerow, J., & Dunemann, L. (2003). Arsen-Metabolismus im Menschen: Kopplung von Festphasenmikroextraktion und GC-MS. *Chemie in unserer Zeit, 37*(4), 248–256. http://dx.doi.org/10.1002/ciuz.200300250
WCZ03_009	Fuhrmann, H., Dwars, T., & Oehme, G. (2003). Wasser als Lösungsmittel: Koordinationskatalyse. *Chemie in unserer Zeit, 37*(1), 40–50.

	http://dx.doi.org/10.1002/ciuz.200390004
WCZ03_010	Deberitz, J., & Boche, G. (2003). Lithium und seine Verbindungen - Industrielle, medizinische und wissenschaftliche Bedeutung.*Chemie in unserer Zeit, 37*(4), 258–266. http://dx.doi.org/10.1002/ciuz.200300264
WCZ98_001	Arduengo, A. J. III, & Krafczyk, R. (1998). Auf der Suche nach stabilen Carbenen. *Chemie in unserer Zeit, 32*(1), 6–14. http://dx.doi.org/10.1002/ciuz.19980320103
WCZ98_002	Plass, W. (1998). Design magnetischer Materialien: Chemie der Magnete. *Chemie in unserer Zeit, 32*(6), 323–333. http://dx.doi.org/10.1002/ciuz.19980320606
WCZ98_003	Geneste, H., & Hesse, M. (1998). Polyamine und Polyamin-Derivate in der Natur. *Chemie in unserer Zeit, 32*(4), 206–218. http://dx.doi.org/10.1002/ciuz.19980320406
WCZ98_004	Wirz, J. (1998). Carbonylverbindungen als Kohlenstoffsäuren. *Chemie in unserer Zeit, 32*(6), 311–322. http://dx.doi.org/10.1002/ciuz.19980320605
WCZ98_005	Paul, D. (1998). Polymermembranen für die Stofftrennung. *Chemie in unserer Zeit, 32*(4), 197–205. http://dx.doi.org/10.1002/ciuz.19980320405
WCZ98_006	Wamhoff, H., Richardt, G., Schneider, V., & Tulke, A. (1998). Wein und Gesundheit. *Chemie in unserer Zeit, 32*(2), 87–93. http://dx.doi.org/10.1002/ciuz.19980320205
WCZ98_007	Binnewies, M. (1998). Chemische Transportreaktionen. *Chemie in unserer Zeit, 32*(1), 15–21. http://dx.doi.org/10.1002/ciuz.19980320104
WCZ98_008	Uhlmann, E. (1998). Antisense-Oligonucleotide - ein universelles Therapieprinzip. *Chemie in unserer Zeit, 32*(3), 150–160. http://dx.doi.org/10.1002/ciuz.19980320306
WCZ98_009	Guthausen, A., Zimmer, G., Laukemper-Ostendorf, S., Blümler, P., & Blümich, B. (1998). NMR-BiIdgebung und MateriaIforschung. *Chemie in unserer Zeit, 32*(2), 73–82. http://dx.doi.org/10.1002/ciuz.19980320204
WCZ98_010	Himmel, H.-J., & Wöll, C. (1998). Herstellung organischer Dünnstschichten. *Chemie in unserer Zeit, 32*(6), 294–301. http://dx.doi.org/10.1002/ciuz.19980320603
WDMW03_001	Hauner, H., Köster, I., & von Ferber, L. (2003). Prävalenz des Diabetes mellitus in Deutschland 1998–2001. Sekundärdatenanalyse einer Versichertenstichprobe der AOK Hessen/KV Hessen. *Deutsche Medizinsche Wochenschrift, 128*(50), 2632–2638. http://dx.doi.org/10.1055/s-2003-812396
WDMW03_002	Baumann, D., Pusterla, N., Péter, O., Grimm, F., Fournier, P. E., Schär, G., Bossart, W., Lutz, H. & Weber, R. (2003). Fieber nach Zeckenstich: Klinik und Diagnostik von akuten Zeckenstich-assoziierten Infektionskrankheiten in der Nordostschweiz. *Deutsche Medizinsche Wochenschrift, 128*(19), 1042–1047. http://dx.doi.org/10.1055/s-2003-39103
WDMW03_003	Hauner, H., Köster, I., & von Ferber, L. (2003). Ambulante Versorgung von Patienten mit Diabetes mellitus im Jahr 2001. Analyse einer Versichertenstichprobe der AOK Hessen/KV Hessen. *Deutsche Medizinsche Wochenschrift, 128*(50), 2638–2643. http://dx.doi.org/10.1055/s-2003-45484
WDMW03_004	Weide, R., Engelhart, S., Färber, H., Kaufmann, F., Heymanns, J., & Köppler, H. (2003). Chronische Bleivergiftung durch ayurvedische Heilpillen. *Deutsche Medizinsche Wochenschrift, 128*(46), 2418–2420. http://dx.doi.org/10.1055/s-2003-43590
WDMW03_005	Schwenger, V., Hofmann, A., Khalifeh, N., Meyer, T., Zeier, M., Hörl, W. H., & Ritz, E. (2003). Urämische Patienten – späte Überweisung, früher Tod. *Deutsche Medizinsche Wochenschrift, 128*(22), 1216–1220.

	http://dx.doi.org/10.1055/s-2003-39471
WDMW03_006	Grimm, W., & Fischbach, W. (2003). Helicobacter pylori-Infektion bei Kindern und Jugendlichen. Eine epidemiologische Untersuchung zu Prävalenz, sozioökonomischen Faktoren und Beschwerdebild. *Deutsche Medizinsche Wochenschrift, 128*(37), 1878–1883. http://dx.doi.org/10.1055/s-2003-42158
WDMW03_007	Lestin, F., Pertschy, A., & Rimek, D. (2003). Fungämie nach oraler Gabe von *Saccharomyces boulardii* bei einem multimorbiden Patienten. *Deutsche Medizinsche Wochenschrift, 128*(48), 2531–2533. http://dx.doi.org/10.1055/s-2003-44948
WDMW03_008	Schulze, J., Rothe, U., Müller, G., Kunath, H., & Fachkommission Diabetes Sachsen (2003). Verbesserung der Versorgung von Diabetikern durch das sächsische Betreuungsmodell. *Deutsche Medizinsche Wochenschrift, 128*(21), 1161–1166. http://dx.doi.org/10.1055/s-2003-39353
WDMW03_009	Küpper-Nybelen, J., Rothenbacher, D., Hahmann, H., Wüsten, B., & Brenner, H. (2003). Veränderungen von Risikofaktoren nach stationärer Rehabilitation bei Patienten mit koronarer Herzkrankheit. *Deutsche Medizinsche Wochenschrift, 128*(28-29), 1525–1530. http://dx.doi.org/10.1055/s-2003-40388
WDMW03_010	Fux, C., Bodmer, T., Ziswiler, H.-R., & Leib, S. L. (2003). *Nocardia cyriacigeorgici*: Erstbeschreibung als invasive Infektion. *Deutsche Medizinsche Wochenschrift, 128*(19), 1038–1041. http://dx.doi.org/10.1055/s-2003-39102
WDMW98_001	Strahl, S., Ehret, V., Dahm, H. H., & Maier, K. P. (1998). Nekrotisierende Hepatitis nach Einnahme pflanzlicher Heilmittel. *Deutsche Medizinsche Wochenschrift, 123*(47), 1410–1414. http://dx.doi.org/10.1055/s-2007-1024196
WDMW98_002	Bergant, A. M., Nguyen, T., Heim, K., Ulmer, H., & Dapunt, O. (1998). Deutschsprachige Fassung und Validierung der »Edinburgh postnatal depression scale«. *Deutsche Medizinsche Wochenschrift, 123*(3), 35–40. http://dx.doi.org/10.1055/s-2007-1023895
WDMW98_003	Reuhl, T., Kaisers, H., Markwardt, J., Haensch, W., Hohenberger, P., & Schlag, P. M. (1998). Axillaausräumung bei klinisch nodal-negativem Mammakarzinom. Kann die Indikation durch »sentinel node«-Nachweis individualisiert werden? *Deutsche Medizinsche Wochenschrift, 123*(19), 583–587. http://dx.doi.org/10.1055/s-2007-1024023
WDMW98_004	Grothey, A., Düppe, J., Hasenburg, A., & Voigtmann, R. (1998). Anwendung alternativmedizinischer Methoden durch onkologische Patienten. *Deutsche Medizinsche Wochenschrift, 123*(31-32), 923–929. http://dx.doi.org/10.1055/s-2007-1024099
WDMW98_005	Schannwell, C. M., Schoebel, F. C., Badiian, M., Jax, T. W., Marx, R., Plehn, G., Perings, C., Vester, E. G., Leschke, M & Strauer, B. E. (1998). Diastolische Funktionsparameter und atriale Rhythmusstörungen bei Patienten mit arterieller Hypertonie. *Deutsche Medizinsche Wochenschrift, 123*(33), 957–964. http://dx.doi.org/10.1055/s-2007-1024104
WDMW98_006	Tsokos, M., Bartel, A., Schoel, R., Rabenhorst, G., & Schwerk, W.-B. (1998). Tödliche Lungenarterienembolie nach endoskopischer Embolisation einer »Downhill-Varize« des Ösophagus. *Deutsche Medizinsche Wochenschrift, 123*(22), 691–695. http://dx.doi.org/10.1055/s-2007-1024039

WDMW98_007	Messmann, H., Knüchel, R., Endlicher, E., Hauser, T., Szeimies, R. M., Kullmann, F., Bäumler, W. & Schölmerich, J. (1998). Photodynamische Diagnostik gastrointestinaler Präkanzerosen nach Sensibilisierung mit 5-Aminolävulinsäure. Eine Pilotstudie. *Deutsche Medizinsche Wochenschrift, 123*(17), 515–521. http://dx.doi.org/10.1055/s-2007-1024003
WDMW98_008	Csef, H., & Heindl, B. (1998). Einstellungen zur Sterbehilfe bei deutschen Ärzten: Eine repräsentative Befragung im Ärztlichen Kreisverband Würzburg. *Deutsche Medizinsche Wochenschrift, 123*(50), 1501–1506. http://dx.doi.org/10.1055/s-2007-1024439
WDMW98_009	Pfaffenbach, B., Götze, O., Szymanski, C., Hagemann, D., & Adamek, R. J. (1998). ^{13}C-Methacetin-Atemtest zur quantitativen nicht-invasiven Leberfunktionsanalyse mittels eines isotopenselektiven nicht-dispersiven Infrarotspektrometers bei Leberzirrhose. *Deutsche Medizinsche Wochenschrift, 123*(49), 1467–1471. http://dx.doi.org/10.1055/s-2007-1024249
WDMW98_010	Hagenah, W., Dörges, I., Gafumbegete, E., & Wagner, T. (1998). Subkutane Manifestationen eines zentrozytischen Non-Hodgkin-Lymphoms an Injektionsstellen eines Mistelpräparats. *Deutsche Medizinsche Wochenschrift, 123*(34-35), 1001–1004. http://dx.doi.org/10.1055/s-2007-1024111
WOrth03_001	Hofmann, S., Romero, J., Roth-Schiffl, E., & Albrecht, T. (2003). Rotationsfehlstellungen der Komponenten als Ursache chronischer Schmerzen und vorzeitigem Prothesenversagen bei Knieendoprothesen. *Der Orthopäde, 32*(6), 469–476. http://dx.doi.org/10.1007/s00132-003-0503-5
WOrth03_002	Thomas, P. (2003). Allergien durch Implantatwerkstoffe. *Der Orthopäde, 32*(1), 60–64. http://dx.doi.org/10.1007/s00132-002-0413-y
WOrth03_003	Romero, J., Stähelin, T., Wyss, T., & Hofmann, S. (2003). Die Bedeutung der axialen Rotationsausrichtung der Knieprothesenkomponenten. *Der Orthopäde, 32*(6), 461–468. http://dx.doi.org/10.1007/s00132-003-0475-5
WOrth03_004	Schnürer, S. M., Gopp, U., Kühn, K.-D., & Breusch, S. J. (2003). Knochenersatzwerkstoffe. *Der Orthopäde, 32*(1), 2–10. http://dx.doi.org/10.1007/s00132-002-0407-9
WOrth03_005	Köck, F. X., Borisch, N., Koester, B., & Grifka, J. (2003). Das komplexe regionale Schmerzsyndrom Typ I (CRPS I). Ursache, Diagnostik und Therapie. *Der Orthopäde, 32*(5), 418–431. http://dx.doi.org/10.1007/s00132-003-0468-4
WOrth03_006	König, A., & Kirschner, S. (2003). Langzeitergebnisse in der Knieendoprothetik. *Der Orthopäde, 32*(6), 516–526. http://dx.doi.org/10.1007/s00132-003-0481-7
WOrth03_007	Beckenbaugh, R. D. (2003). Die Arthroplastik des Metakarpophalangealgelenkes mit Pyrocarbonimplantaten. *Der Orthopäde, 32*(9), 794–797. http://dx.doi.org/10.1007/s00132-003-0519-x
WOrth03_008	Bernau, A., & Heeg, P. (2003). Intraartikuläre Punktionen und Injektionen. *Der Orthopäde, 32*(6), 548–570. http://dx.doi.org/10.1007/s00132-003-0498-y
WOrth03_009	Gerdesmeyer, L., Rechl, H., Wagenpfeil, S., Ulmer, M., Lampe, R., & Wagner, K. (2003). Die minimal-invasive perkutane epidurale Neurolyse beim chronischen Nervenwurzelreizsyndrom. Eine prospektive kontrollierte Pilotstudie zum Wirksamkeitsnachweis. *Der Orthopäde, 32*(10), 869–876. http://dx.doi.org/10.1007/s00132-003-0533-z

WOrth03_010	Sparmann, M., & Wolke, B. (2003). Stellenwert der Navigation und Roboterchirurgie bei Knie-Totalendoprothesen. *Der Orthopäde, 32*(6), 498–505. http://dx.doi.org/10.1007/s00132-003-0479-1
WOrth98_001	Rueger, J. M. (1998). Knochenersatzmittel. Heutiger Stand und Ausblick. *Der Orthopäde, 27*(2), 72–79. http://dx.doi.org/10.1007/PL00003481
WOrth98_002	Leunig, M., & Ganz, R. (1998). Berner periazetabuläre Osteotomie. *Der Orthopäde, 27*(11), 743–750. http://dx.doi.org/10.1007/PL00003460
WOrth98_003	Günther, K. P., Scharf, H.-P., Pesch, H.-J., & Puhl, W. (1998). Einwachsverhalten von Knochenersatzstoffen. Tierexperimentelle Untersuchung. *Der Orthopäde, 27*(2), 105–117. http://dx.doi.org/10.1007/PL00003476
WOrth98_004	Rueger, J. M., Linhart, W., & Sommerfeldt, D. (1998). Biologische Reaktionen auf Kalziumphosphatkeramik-Implantationen. Tierexperimentelle Ergebnisse. *Der Orthopäde, 27*(2), 89–95. http://dx.doi.org/10.1007/PL00003483
WOrth98_005	Millis, M. B., & Murphy, S. B. (1998). Das Bostoner Konzept: Die periazetabuläre Osteotomie mit simultaner Arthrotomie über den direkten vorderen Zugang. *Der Orthopäde, 27*(11), 751–758. http://dx.doi.org/10.1007/PL00003461
WOrth98_006	Hofmann, S., Tschauner, C., Urban, M., Eder, T., & Czerny, C. (1998). Klinische und bildgebende Diagnostik der Labrumläsion des Hüftgelenks. *Der Orthopäde, 27*(10), 681–689. http://dx.doi.org/10.1007/PL00003453
WOrth98_007	Kerschbaumer, F., Kandziora, F., Herresthal, J., Hertel, A., & Hör, G. (1998). Synovektomie und Synoviorthese als Kombinationstherapie bei rheumatoider Arthritis. *Der Orthopäde, 27*(3), 188–196. http://dx.doi.org/10.1007/PL00003490
WOrth98_008	Blauth, M., Knop, C., Bastian, L., Krettek, C., & Lange, U. (1998). Komplexe Verletzungen der Wirbelsäule. *Der Orthopäde, 27*(1), 17–31. http://dx.doi.org/10.1007/PL00003446
WOrth98_009	Hedtmann, A., Fett, H., & Ludwig, J. (1998). Die Behandlung veralteter, posttraumatischer Akromioklavikulargelenkinstabilitäten und -arthrosen. *Der Orthopäde, 27*(8), 556–566. http://dx.doi.org/10.1007/PL00003528
WOrth98_010	Putz, R., & Schrank, C. (1998). Anatomie des labrokapsulären Komplexes. *Der Orthopäde, 27*(10), 675–680. http://dx.doi.org/10.1007/PL00003452
ZPad03_001	Konietzka, D., & Seibert, H. (2003). Deutsche und Ausländer an der "zweiten Schwelle". Eine vergleichende Analyse der Berufseinstiegskohorten 1976-1995 in Westdeutschland. *Zeitschrift für Pädagogik, 49*(4), 567–590. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-38936
ZPad03_002	Fuchs, H.-W. (2003). Auf dem Weg zu einem Weltcurriculum? Zum Grundbildungskonzept von PISA und der Aufgabenzuweisung an die Schule. *Zeitschrift für Pädagogik, 49*(2), 161–179. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-38727
ZPad03_003	Bynner, J., Schuller, T., & Feinstein, L. (2003). Wider benefits of education: skills, higher education and civic engagement. *Zeitschrift für Pädagogik, 49*(3), 341–361. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-38821
ZPad03_004	Bos, W., Lankes, E.-M., Prenzel, M., Schwippert, K., Walther, G., Valtin, R., & Voss, A. (2003). Welche Fragen können aus einer gemeinsamen Interpretation der Befunde aus PISA und IGLU fundiert beantwortet werden? *Zeitschrift für Pädagogik, 49*(2), 198–212. Retrieved September

	16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-38741
ZPad03_005	Oesterreich, D. (2003). Offenes Diskussionsklima im Unterricht und politische Bildung von Jugendlichen. *Zeitschrift für Pädagogik, 49*(6), 817–836. Retrieved from http://nbn-resolving.de/urn:nbn:de:0111-opus-39051
ZPad03_006	Depaepe, M., & Simon, F. (2003). Freiluftschulen: eine historisch-pädagogische Randerscheinung als Reflex sozial-historischer Modernisierungsprozesse? Das Beispiel Belgiens. *Zeitschrift für Pädagogik, 49*(5), 718–733. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-39004
ZPad03_007	Wild, E. (2003). Einbeziehung des Elternhauses durch Lehrer: Art, Ausmaß und Bedingungen der Elternpartizipation aus der Sicht von Gymnasiallehrern. *Zeitschrift für Pädagogik, 49*(4), 513–533. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-38907
ZPad03_008	Alheit, P. (2003). Mentalität und Intergenerationalität als Rahmenbedingungen "Lebenslangen Lernens". Konzeptionelle Konsequenzen aus Ergebnissen einer biografieanalytischen Mehrgenerationenstudie in Ostdeutschland. *Zeitschrift für Pädagogik, 49*(3), 362–382. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-38836
ZPad03_009	Langewand, A. (2003). Über die Schwierigkeit, Erziehung als Aufforderung zur Selbsttätigkeit zu begreifen. *Zeitschrift für Pädagogik, 49*(2), 274–289. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-38780
ZPad03_010	Roeder, P. M. (2003). TIMSS und PISA - Chancen eines neuen Anfangs in Bildungspolitik, -planung, -verwaltung und Unterricht. Endlich ein Schock mit Folgen? *Zeitschrift für Pädagogik, 49*(2), 180–197. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-38736
ZPad08_001	Klieme, E., & Rakoczy, K. (2008). Empirische Unterrichtsforschung und Fachdidaktik. Outcome-orientierte Messung und Prozessqualität des Unterrichts. *Zeitschrift für Pädagogik, 54*(2), 222–237. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-43488
ZPad08_002	Maier, U. (2008). Rezeption und Nutzung von Vergleichsarbeiten aus der Perspektive von Lehrkräften. *Zeitschrift für Pädagogik, 54*(1), 95–117. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-43384
ZPad08_003	Helsper, W. (2008). Schulkulturen - die Schule als symbolische Sinnordnung. *Zeitschrift für Pädagogik, 54*(1), 63–80. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-43365
ZPad08_004	Koretz, D. (2008). Test-Based Educational Accountability. Research Evidence and Implications. *Zeitschrift für Pädagogik, 54*(6), 777–790. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-43768
ZPad08_005	Petermann, F., & Natzke, H. (2008). Aggressives Verhalten in der Schule. Ausdrucksformen, Verlaufsmuster und Möglichkeiten entwicklungsorientierter Prävention. *Zeitschrift für Pädagogik, 54*(4), 532–554. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-43635
ZPad08_006	Pant, H. A., Vock, M., Pöhlmann, C., & Köller, O. (2008). Offenheit für Innovationen. Befunde aus einer Studie zur Rezeption der Bildungsstandards bei Lehrkräften und Zusammenhänge mit Schülerleistungen. *Zeitschrift für Pädagogik, 54*(6), 827–845. Retrieved

	September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-43796
ZPad08_007	Reyer, J., & Franke-Meyer, D. (2008). Muss der Bildungsauftrag des Kindergartens "eigenständig" sein? *Zeitschrift für Pädagogik, 54*(6), 888–905. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-43836
ZPad08_008	Boreham, N., & Reeves, J. (2008). Diagnosing and Supporting a Culture of Organizational Learning in Scottish schools. *Zeitschrift für Pädagogik, 54*(5), 637–649. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-43688
ZPad08_009	Stojanov, K. (2008). Bildungsgerechtigkeit als Freiheitseinschränkung? Kritische Anmerkungen zum Gebrauch der Gerechtigkeitskategorie in der empirischen Bildungsforschung. *Zeitschrift für Pädagogik, 54*(4), 516–531. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-43620
ZPad08_010	Schreiber, W. (2008). Ein Kompetenz-Strukturmodell historischen Denkens. *Zeitschrift für Pädagogik, 54*(2), 198–212. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-43457
ZPad98_001	Nauck, B., Diefenbach, H., & Petri, K. (1998). Intergenerationale Transmission von kulturellem Kapital unter Migrationsbedingungen. Zum Bildungserfolg von Kindern und Jugendlichen aus Migrantenfamilien in Deutschland. *Zeitschrift für Pädagogik, 44*(5), 701–722. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-68364
ZPad98_002	Bauer, K.-O. (1998). Pädagogisches Handlungsrepertoire und professionelles Selbst von Lehrerinnen und Lehrern. *Zeitschrift für Pädagogik, 44*(3), 343–359. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-68216
ZPad98_003	Biskup, C., Pfister, G., & Röbke, C. (1998). "Weil man da über seine Probleme reden kann…". Partielle Geschlechtertrennung aus der Sicht der Schülerinnen und Schüler. *Zeitschrift für Pädagogik, 44*(5), 753–768. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-68390
ZPad98_004	Hagemann, W., & Rose, F.-J. (1998). Zur Lehrer/innen-Erfahrung von Lehramts-Studierenden. *Zeitschrift für Pädagogik, 44*(1), 7–19. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-68020
ZPad98_005	Harteis, C., & Prenzel, M. (1998). Welche Kompetenzen brauchen betriebliche Weiterbildner in Zukunft? Ergebnisse einer Delphi-Studie in einem Industrieunternehmen. *Zeitschrift für Pädagogik, 44*(4), 583–601. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-68316
ZPad98_006	Zymek, B. (1998). "Leitbild ist nicht mehr der erwerbstätige, sondern der tätige Mensch". Ein bildungshistorischer Kommentar zu den Forderungen der Kommission für Zukunftsfragen der Freistaaten Bayern und Sachsen. *Zeitschrift für Pädagogik, 44*(6), 789–803. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-68417
ZPad98_007	Gogolin, I., Neumann, U., & Reuter, L. (1998). Schulbildung für Minderheiten. Eine Bestandsaufnahme. *Zeitschrift für Pädagogik, 44*(5), 663–678. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-68341

ZPad98_008	Kluchert, G., & Leschinsky, A. (1998). Schule in der Transformation - Transformation der Schule? Was man aus Gesprächen mit ehemaligen Schülern über die Schule "zwischen zwei Diktaturen" erfahren kann. *Zeitschrift für Pädagogik, 44*(4), 543–564. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-68299
ZPad98_009	Krause, G., & Wenzel, H., u.a. (1998). Lehrerbewußtsein und Handlungsstrukturen im Wendeprozeß. *Zeitschrift für Pädagogik, 44*(4), 565–581. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-68301
ZPad98_010	Beetz, S. (1998). Koedukationsdiskurs zwischen Programmatik und Erfahrungswissen. Von der Notwendigkeit einer Inszenierung entdramatisierter Geschlechterverhältnisse im Bildungswesen. *Zeitschrift für Pädagogik, 44*(2), 253–262. Retrieved September 16, 2014 from http://nbn-resolving.de/urn:nbn:de:0111-opus-68163

C Cited Reference Search

Appendix C describes the methodology of the *Cited Reference Search* as applied to identify the missed citations for the cited articles of the data sample in WoS.

As described in section 5.6 we divided the data sample into 12 different datasets according to their journal name. Each dataset was separately searched in the *Cited Reference Search*. We always started off by searching for variations in the publication names. First, we included the comment *in press*, for which the publication year would be missing in the reference. Afterwards, we looked for variations of the publication name itself using as few letters and as many wildcards as possible (cf. Table 37, row 5). Depending on the number of results, we would try other variations including more letters (rows 6-12). In this example, we did not find many variations of the publication name and, therefore, switched to looking for permutations of author names. In other cases, where the permutations of publications retrieved more results, we either looked through the result set directly (up to 200 result records) or combined the variation with either the publication year or the author name.

Analogously to the permutations of the publication name, the second step in the search was to look for variations of the author name. Again, we started by looking for name variations with the fewest letters possible. Depending on the number of results we either looked through the records directly or combined these with variations of the publication name and then with the publication year. In order to find as many missed citations as possible, ideally all missed citations to a cited article, we then varied possible error sources of the author name, such as vowels, phonetic similar letters, neighboring consonants or switching first and last name as well as trying different permutations of compounded names, as described in section 4.2 on inaccuracies in bibliographic data values. Again, these were either browsed through directly or combined with variations of publication names and/or publication years. Even though, the *Cited Reference Search* in WoS offers the possibility not only to look for variations of publication name, author name and publication year, but also for variations of the volume number, issue, page numbers and even article title, we did not employ those additional

bibliographic fields in our search strategy. This decision was taken in the course of the search, because we were able to obtain a manageable number of result records from the combination of publication name, author name and publication year. WoS also points out on their *Cited Reference Search* page, that "Entering the title, volume, issue, or page in combination with other fields may reduce the number of citing reference variants found".

Table 37: Example of publication name variations of Political Theory[51]

Row	Permutation	No of results
1	pol* the* in press	1
2	pol* the* inpress	0
3	in press pol* the*	4
4	inpress pol* the*	0
5	p*l t*y	17,084
6	plt* theo*	0
7	plot* theo*	5
8	polt*l theo*	15
9	polt* the*	24
10	polical the*	0
11	politcal th*	5
12	polital th*	0

The verification, whether the citing article truly contained a citation to the cited article, was carried out in the data entry process by the two student assistants. In case they did not find the citation in the article, we double-checked them. Only in a few cases, the missed citation we found in the *Cited Reference Search* did not point to the correct target article. Most of them cited another article in the same journal but in a different year or with a different issue number.

[51] The number of records was last checked August 1, 2014 in the *Cited Reference Search* of WoS.

D DATA PARSING PROCEDURES

Table 38: Non-alphanumeric characters that were eliminated from the article title, publication name and author name

.	´	“	"	”
’	`	'	“	′
:	’	'	″	”
;	‘	*	+	–
—	-	/	( )	[ ]
« »	<>	&	?	!

List of stop words identified in the assessment process

- the
- a
- and
- ein
- on
- of
- in
- to

Code of Levenshtein distance function

(Retrieved September 18, 2014 from http://stackoverflow.com/questions/13909885/how-to-add-levenshtein-function-in-mysql):

```
DELIMITER $$

CREATE FUNCTION levenshtein( s1 VARCHAR(255), s2 VARCHAR(255) )

  RETURNS INT
```

```
      DETERMINISTIC
      BEGIN
        DECLARE s1_len, s2_len, i, j, c, c_temp, cost INT;
        DECLARE s1_char CHAR;
        -- max strlen=255
        DECLARE cv0, cv1 VARBINARY(256);
        SET s1_len = CHAR_LENGTH(s1), s2_len = CHAR_LENGTH(s2), cv1 = 0x00, j = 1, i =
1, c = 0;
        IF s1 = s2 THEN
          RETURN 0;
        ELSEIF s1_len = 0 THEN
          RETURN s2_len;
        ELSEIF s2_len = 0 THEN
          RETURN s1_len;
        ELSE
          WHILE j <= s2_len DO
            SET cv1 = CONCAT(cv1, UNHEX(HEX(j))), j = j + 1;
          END WHILE;
          WHILE i <= s1_len DO
            SET s1_char = SUBSTRING(s1, i, 1), c = i, cv0 = UNHEX(HEX(i)), j = 1;
            WHILE j <= s2_len DO
              SET c = c + 1;
              IF s1_char = SUBSTRING(s2, j, 1) THEN
                SET cost = 0; ELSE SET cost = 1;
              END IF;
```

```
        SET c_temp = CONV(HEX(SUBSTRING(cv1, j, 1)), 16, 10) + cost;
        IF c > c_temp THEN SET c = c_temp; END IF;
          SET c_temp = CONV(HEX(SUBSTRING(cv1, j+1, 1)), 16, 10) + 1;
          IF c > c_temp THEN
            SET c = c_temp;
          END IF;
          SET cv0 = CONCAT(cv0, UNHEX(HEX(c))), j = j + 1;
      END WHILE;
      SET cv1 = cv0, i = i + 1;
    END WHILE;
  END IF;
  RETURN c;
END$$
```

Table 39: List of special characters that were tested with the LDF

ID	Special_character	Equiv_character	LDF_score	Subset MS Office
1	à	a	0	Latin-1_Suppl
2	á	a	0	Latin-1_Suppl
3	â	a	0	Latin-1_Suppl
4	ã	a	0	Latin-1_Suppl
5	ä	a	1	Latin-1_Suppl
6	å	a	1	Latin-1_Suppl
7	æ	a	1	Latin-1_Suppl
8	ç	c	0	Latin-1_Suppl
9	è	e	0	Latin-1_Suppl
10	é	e	0	Latin-1_Suppl
11	ê	e	0	Latin-1_Suppl
12	ë	e	0	Latin-1_Suppl
13	ì	i	0	Latin-1_Suppl
14	í	i	0	Latin-1_Suppl
15	î	i	0	Latin-1_Suppl
16	ï	i	0	Latin-1_Suppl
17	ñ	n	0	Latin-1_Suppl
18	ò	o	0	Latin-1_Suppl

ID	Special_character	Equiv_character	LDF_score	Subset MS Office
19	ó	o	0	Latin-1_Suppl
20	ô	o	0	Latin-1_Suppl
21	õ	o	0	Latin-1_Suppl
22	ö	o	1	Latin-1_Suppl
23	ø	o	1	Latin-1_Suppl
24	ù	u	0	Latin-1_Suppl
25	ú	u	0	Latin-1_Suppl
26	û	u	0	Latin-1_Suppl
27	ü	u	1	Latin-1_Suppl
28	ý	y	0	Latin-1_Suppl
29	ÿ	y	1	Latin-1_Suppl
30	ā	a	1	Latin-Ext-A
31	ă	a	1	Latin-Ext-A
32	ą	a	1	Latin-Ext-A
33	ć	c	1	Latin-Ext-A
34	ĉ	c	1	Latin-Ext-A
35	ċ	c	1	Latin-Ext-A
36	č	c	1	Latin-Ext-A
37	ď	d	1	Latin-Ext-A
38	đ	d	1	Latin-Ext-A
39	ē	e	1	Latin-Ext-A
40	ĕ	e	1	Latin-Ext-A
41	ė	e	1	Latin-Ext-A
42	ę	e	1	Latin-Ext-A
43	ě	e	1	Latin-Ext-A
44	ĝ	g	1	Latin-Ext-A
45	ğ	g	1	Latin-Ext-A
46	ġ	g	1	Latin-Ext-A
47	ģ	g	1	Latin-Ext-A
48	ĥ	h	1	Latin-Ext-A
49	ħ	h	1	Latin-Ext-A
50	ĩ	i	1	Latin-Ext-A
51	ī	i	1	Latin-Ext-A
52	ĭ	i	1	Latin-Ext-A
53	į	i	1	Latin-Ext-A
54	ı	i	1	Latin-Ext-A
55	ĵ	j	1	Latin-Ext-A
56	ķ	k	1	Latin-Ext-A
57	Ĺ	l	1	Latin-Ext-A
58	ļ	l	1	Latin-Ext-A
59	ľ	l	1	Latin-Ext-A
60	ŀ	l	1	Latin-Ext-A
61	ł	l	1	Latin-Ext-A
62	ń	n	1	Latin-Ext-A

ID	Special_character	Equiv_character	LDF_score	Subset MS Office
63	ņ	n	1	Latin-Ext-A
64	ň	n	1	Latin-Ext-A
65	ʼn	n	1	Latin-Ext-A
66	ŋ	n	1	Latin-Ext-A
67	ō	o	1	Latin-Ext-A
68	ŏ	o	1	Latin-Ext-A
69	ő	o	1	Latin-Ext-A
70	œ	o	1	Latin-Ext-A
71	ŕ	r	1	Latin-Ext-A
72	ŗ	r	1	Latin-Ext-A
73	ř	r	1	Latin-Ext-A
74	ś	s	1	Latin-Ext-A
75	ŝ	s	1	Latin-Ext-A
76	ş	s	1	Latin-Ext-A
77	š	s	1	Latin-Ext-A
78	ţ	t	1	Latin-Ext-A
79	ť	t	1	Latin-Ext-A
80	ŧ	t	1	Latin-Ext-A
81	ũ	u	1	Latin-Ext-A
82	ū	u	1	Latin-Ext-A
83	ŭ	u	1	Latin-Ext-A
84	ů	u	1	Latin-Ext-A
85	ű	u	1	Latin-Ext-A
86	ų	u	1	Latin-Ext-A
87	ŵ	w	1	Latin-Ext-A
88	ŷ	y	1	Latin-Ext-A
89	ź	z	1	Latin-Ext-A
90	ż	z	1	Latin-Ext-A
91	ž	z	1	Latin-Ext-A

Table 40: Special characters that the LDF cannot detect

ID	Special_character	Equiv_character	LDF_score	Subset MS Office
1	à	a	0	Latin-1_Suppl
2	á	a	0	Latin-1_Suppl
3	â	a	0	Latin-1_Suppl
4	ã	a	0	Latin-1_Suppl
8	ç	c	0	Latin-1_Suppl
9	è	e	0	Latin-1_Suppl
10	é	e	0	Latin-1_Suppl
11	ê	e	0	Latin-1_Suppl
12	ë	e	0	Latin-1_Suppl
13	ì	i	0	Latin-1_Suppl
14	í	i	0	Latin-1_Suppl
15	î	i	0	Latin-1_Suppl
16	ï	i	0	Latin-1_Suppl
17	ñ	n	0	Latin-1_Suppl
18	ò	o	0	Latin-1_Suppl
19	ó	o	0	Latin-1_Suppl
20	ô	o	0	Latin-1_Suppl
21	õ	o	0	Latin-1_Suppl
24	ù	u	0	Latin-1_Suppl
25	ú	u	0	Latin-1_Suppl
26	û	u	0	Latin-1_Suppl
28	ý	y	0	Latin-1_Suppl

E THE CODEBOOK

Table 41: The codebook

Inaccuracy code	Name	Type 1: contains a correct value	Type 2: contains part of a correct value	Type 3: does not contain a correct value
A	Typographical variation		x	
B	Spelling error		x	
C	Different language			x
D	Completely incorrect			x
E	Omitted			x
F	Cropped		x	
G	Interchanged fields	x		
G1	holds issue no	x		
G2	holds starting page	x		
G3	holds ending page	x		
G4	holds volume no	x		
G5	holds last name	x		
G6	holds first initial	x		
G7	holds second initial	x		
H	Jumbled value	x		
I	Abbreviation		x	
J	Partially incorrect		x	
K	Space	x		
L	Informational letter	x		
M	Incorrect interpretation of author names		x	
N	Additional information	x		
O	Incorrect order of authors	x		
P	No author name			x
Q	Special character		x	
R	Punctuation	x		

Inaccuracy code	Name	Type 1: contains a correct value	Type 2: contains part of a correct value	Type 3: does not contain a correct value
S	Padded	x		
T	Plus/Minus		x	
U	Full first name	x		
V	Incorrect interpretation of additional information		x	
X	Stop word		x	
Y	Word stem		x	
Z	Not available			x

F Results of the quantitative analysis

Table 42 summarizes the frequency of IACs in the Orig-WoS result set. The first few rows give the descriptive statistics. Below, the first column names the IAC. The second column shows the absolute number of occurrences of the IAC (*Count*) and third column gives the percentage of these occurrences in the respective inaccuracy category *simple*, *moderate* and *complex* (*Type %*). The IACs are organized according to the taxonomy explained in section 6.3. The last row gives the total absolute number of inaccuracies.

Table 42: Overall frequency of IACs in the Orig-WoS result set

Assessment result Orig-WoS		
No of articles		300
No of assessed data values		4,005
No of inaccuracies		437
% of data values without discrepancy		90%
IAC	Count	Type %
simple		
Added data values		
N *Additional information*	1	8%
S *Padded*	1	8%
Disarranged data values		
G *Interchanged fields*	4	34%
O *Incorrect order of authors*	6	50%
moderate		
Incorrect interpretation of data values		
M *Incorrect interpretation of author names*	10	5%
V *Incorrect interpretation of add. information*	3	2%
Spelling variations		
B *Spelling error*	4	2%
Q *Special character*	114	61%
Abbreviated data values		
F *Cropped*	2	1%

IAC	Count	Type %
I *Abbreviation*	30	16%
Other variations		
T *Plus/Minus*	3	2%
X *Stop word*	20	11%
complex		
Not assessable		
C *Different language*	146	61%
Z *Not available*	20	8%
Missing data values		
E *Omitted*	67	28%
Completely incorrect		
D *Completely incorrect*	6	3%
SUM	437	

Table 43 summarizes the frequency of IACs in the Orig-Ref and WoS-Ref result sets. The first few rows give the descriptive statistics. Below, the first column names the IAC. The following two columns show the results of the Orig-Ref and the last two the results of the WoS-Ref assessment. Each result set gives the absolute number of occurrences of the IAC (*Count*) and the percentage of these occurrences in the respective inaccuracy category *simple*, *moderate* or *complex* (*Type %*). To facilitate comparison between the assessment results and the discussion of the subcategories, the IACs are organized according to the taxonomy. The last row gives the total absolute number of inaccuracies.

Table 43: Overall frequency of IACs in the two assessment samples: Orig-Ref and WoS-Ref

	Orig-Ref		**WoS-Ref**	
No of articles	3,735		3,735	
No of citing references	3,929		3,929	
No of assessed data values	54,828		54,861	
No of inaccuracies	8,108		8,175	
% of data values without discrepancy	85%		85%	
	Orig-Ref		**WoS-Ref**	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	16	3%	15	2%
L *Informational letter*	48	9%	46	5%
N *Additional information*	38	7%	261	29%
S *Padded*	54	10%	101	11%

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
Disarranged data values				
G *Interchanged fields*	84	16%	95	10%
H *Jumbled value*	31	6%	27	3%
O *Incorrect order of authors*	263	49%	369	40%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	77	2%	144	4%
V *Incorrect interpretation of add. information*	5	0%	3	0%
Spelling variations				
A *Typographical variation*	55	1%	48	1%
B *Spelling error*	265	5%	309	9%
Q *Special character*	300	6%	822	23%
Y *Word stem*	38	1%	84	2%
Abbreviated data values				
F *Cropped*	1,647	35%	1,425	39%
I *Abbreviation*	2,170	45%	359	10%
Other variations				
J *Partially incorrect*	54	1%	228	6%
T *Plus/Minus*	166	3%	127	4%
X *Stop word*	41	1%	90	2%
complex				
Not assessable				
C *Different language*	549	20%	787	22%
Z *Not available*	84	3%	669	18%
Missing data values				
E *Omitted*	1,877	68%	1,827	50%
P *No author name*	23	1%	23	1%
Completely incorrect				
D *Completely incorrect*	223	8%	316	9%
SUM	8,108		8,175	

Table 44 and Table 45 show the occurrences of IACs per bibliographic field for the Orig-Ref and the WoS-Ref result.

Table 44: Occurrences of IACs per bibliographic field – Orig-Ref result

Orig-Ref								
IAC	First AN	Other ANs	Article title	Pubname	Pubyear	Volume no	Spage	Epage
simple								
Added data values								
K *Space*	x	x	x	x				
L *Inform. letter*					x		x	x
N *Additional information*			x	x				
S *Padded*			x	x		x	x	x
Disarranged data values								
G *Interch. fields*	x	x				x	x	x
H *Jumbled value*	x	x	x				x	x
O *Incorrect order of authors*	x	x						
moderate								
Incorrect interpretation of data values								
M *Incorrect interpr. of author names*	x	x						
V *Incorrect interpr. of add. information*			x					
Spelling variations								
A *Typogr. variation*			x					
B *Spelling error*	x	x	x	x				
Q *Special character*	x	x	x	x		x		
Y *Word stem*			x	x				
Abbreviated data values								
F *Cropped*	x	x	x	x		x		x
I *Abbrev.*			x	x				
Other variations								
J *Partially incorrect*			x	x				
T *Plus/Minus*					x	x	x	x
X *Stop word*		x	x	x				
complex								
Not assessable								
C *Different language*			x					
Z *Not available*						x		

IAC	First AN	Other ANs	Article title	Pubname	Pubyear	Volume no	Spage	Epage
Missing data values								
E *Omitted*	x	x	x		x	x	x	x
P *No author name*	x	x						
Completely incorrect								
D *Completely incorrect*	x	x		x	x	x	x	x

Table 45: Occurrences of IAC per bibliographic field – WoS-Ref result

WoS-Ref								
IAC	First AN	Other ANs	Article title	Pubname	Pubyear	Volume no	Spage	Epage
simple								
Added data values								
K *Space*	x	x	x	x				
L *Inform. letter*					x		x	
N *Additional information*	x		x	x				
S *Padded*			x	x		x	x	x
Disarranged data values								
G *Interch. fields*	x	x				x	x	x
H *Jumbled value*	x	x	x	x			x	x
O *Incorrect order of authors*	x	x						
moderate								
Incorrect interpretation of data values								
M *Incorrect interpr. of author names*	x	x						
V *Incorrect interpr. of add. information*			x					
Spelling variations								
A *Typogr. variation*			x					
B *Spelling error*	x	x	x	x				
Q *Special character*	x	x	x	x		x		
Y *Word stem*			x	x				
Abbreviated data values								
F *Cropped*	x	x	x	x		x		x
I *Abbrev.*				x				

IAC	First AN	Other ANs	Article title	Pubname	Pubyear	Volume no	Spage	Epage
Other variations								
J *Partially incorrect*			x	x				
T *Plus/Minus*					x	x	x	x
X *Stop word*			x	x				
complex								
Not assessable								
C *Different language*			x					
Z *Not available*								x
Missing data values								
E *Omitted*	x	x	x		x	x	x	x
P *No author name*	x	x						
Completely incorrect								
D *Completely incorrect*	x	x	x	x	x	x	x	x

The following tables (Table 46-Table 76) give the descriptive statistics as well as the frequencies of IACs for the different strata of the data sample ordered by their incidence in chapter 7.

Table 46: Overall descriptive statistics –NS, SSH

	Assessment result Orig-Ref		Assessment result WoS-Ref	
	NS	SSH	NS	SSH
No of citing references	2,496	1,433	2,496	1,433
No of assessed data values	39,684	15,144	39,717	15,144
No of inaccuracies	6,744	1,364	6,163	2,012

Table 47: Frequency of IACs – NS

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	8	2%	10	1%
L *Informational letter*	4	1%	2	0%
N *Additional information*	32	8%	255	34%
S *Padded*	28	7%	81	10%
Disarranged data values				
G *Interchanged fields*	47	11%	50	6%
H *Jumbled value*	27	7%	24	3%
O *Incorrect order of authors*	255	64%	361	46%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	72	2%	139	5%
V *Incorrect interpretation of add. information*	5	0%	3	0%
Spelling variations				
A *Typographical variation*	44	1%	47	2%
B *Spelling error*	184	5%	253	9%
Q *Special character*	280	7%	463	17%
Y *Word stem*	6	0%	58	2%
Abbreviated data values				
F *Cropped*	1,275	32%	1,130	41%
I *Abbreviation*	2,023	50%	327	12%
Other variations				
J *Partially incorrect*	30	1%	193	7%
T *Plus/Minus*	94	2%	62	2%
X *Stop word*	18	0%	70	3%
complex				
Not assessable				
C *Different language*	538	23%	387	15%
Z *Not available*	0	0%	484	18%
Missing data values				
E *Omitted*	1,643	71%	1,600	61%
P *No author name*	5	0%	5	0%
Completely incorrect				
D *Completely incorrect*	126	6%	159	6%
SUM	6,744		6,163	

Table 48: Frequency of IACs – SSH

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	8	6%	5	4%
L *Informational letter*	44	33%	44	34%
N *Additional information*	6	5%	6	5%
S *Padded*	26	19%	20	15%
Disarranged data values				
G *Interchanged fields*	37	28%	45	34%
H *Jumbled value*	4	3%	3	2%
O *Incorrect order of authors*	8	6%	8	6%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	5	1%	5	1%
Spelling variations				
A *Typographical variation*	11	1%	1	0%
B *Spelling error*	81	10%	56	6%
Q *Special character*	20	3%	359	40%
Y *Word stem*	32	4%	26	3%
Abbreviated data values				
F *Cropped*	372	47%	295	33%
I *Abbreviation*	147	19%	32	4%
Other variations				
J *Partially incorrect*	24	3%	35	4%
T *Plus/Minus*	72	9%	65	7%
X *Stop word*	23	3%	20	2%
complex				
Not assessable				
C *Different language*	11	2%	400	40%
Z *Not available*	84	19%	185	19%
Missing data values				
E *Omitted*	234	53%	227	23%
P *No author name*	18	4%	18	2%
Completely incorrect				
D *Completely incorrect*	97	22%	157	16%
SUM	1,364		2,012	

Table 49: Overall descriptive statistics – disciplines

		No. of cited ref.	No. of data values	No of inacc.
Assessment result Orig-Ref	Chem	753	11,148	2,218
	Ortho	836	10,944	1,728
	GenMed	907	17,592	2,798
	EduSci	434	4,896	319
	PolSci	531	5,232	483
	Sociol	468	5,016	562
Assessment result WoS-Ref	Chem	753	11,202	2,207
	Ortho	836	10,944	1,589
	GenMed	907	17,571	2,367
	EduSci	434	4,896	572
	PolSci	531	5,232	769
	Sociol	468	5,016	671

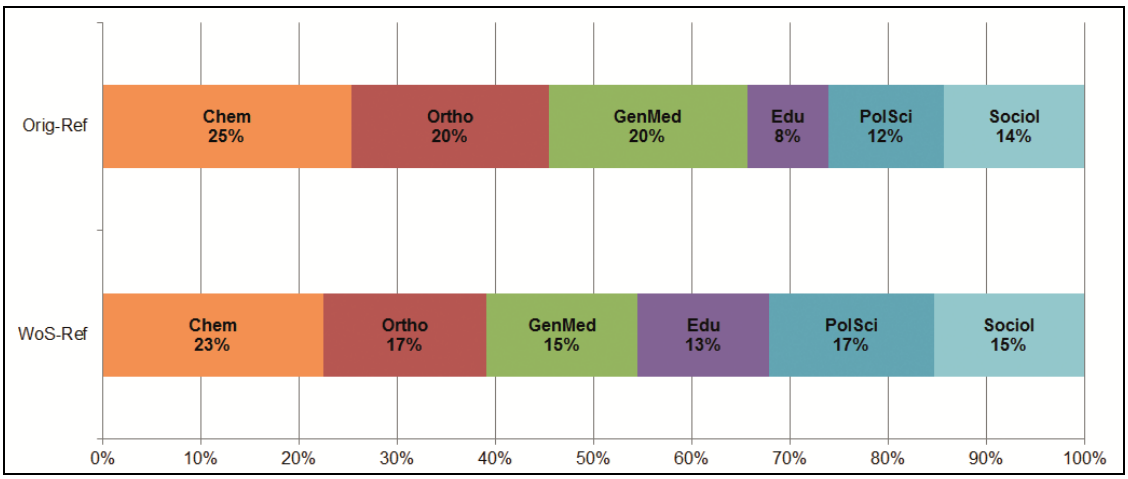


Figure 32: Shares of inaccuracies per discipline

Table 50: Frequency of IACs – Chemistry

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	3	3%	3	1%
L *Informational letter*	0	0%	0	0%
N *Additional information*	4	3%	136	39%
S *Padded*	1	1%	5	1%
Disarranged data values				
G *Interchanged fields*	25	22%	25	7%
H *Jumbled value*	8	7%	4	1%
O *Incorrect order of authors*	74	64%	181	51%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	47	6%	98	18%
V *Incorrect interpretation of add. information*	0	0%	0	0%
Spelling variations				
A *Typographical variation*	0	0%	0	0%
B *Spelling error*	65	8%	74	13%
Q *Special character*	9	1%	57	10%
Y *Word stem*	1	0%	0	0%
Abbreviated data values				
F *Cropped*	71	9%	20	4%
I *Abbreviation*	629	75%	259	46%
Other variations				
J *Partially incorrect*	2	0%	36	7%
T *Plus/Minus*	8	1%	8	1%
X *Stop word*	3	0%	6	1%
complex				
Not assessable				
C *Different language*	13	1%	28	2%
Z *Not available*	0	0%	0	0%
Missing data values				
E *Omitted*	1,218	96%	1,254	97%
P *No author name*	5	0%	5	0%
Completely incorrect				
D *Completely incorrect*	32	3%	8	1%
SUM	2,218		2,207	

Table 51: Frequency of IACs – Orthopedics

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	2	4%	2	2%
L *Informational letter*	2	4%	0	0%
N *Additional information*	2	4%	58	54%
S *Padded*	15	28%	16	15%
Disarranged data values				
G *Interchanged fields*	13	24%	12	11%
H *Jumbled value*	0	0%	1	1%
O *Incorrect order of authors*	19	36%	18	17%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	2	0%	2	0%
V *Incorrect interpretation of add. information*	0	0%	0	0%
Spelling variations				
A *Typographical variation*	9	1%	1	0%
B *Spelling error*	34	3%	68	10%
Q *Special character*	102	8%	272	40%
Y *Word stem*	3	0%	16	2%
Abbreviated data values				
F *Cropped*	338	26%	206	31%
I *Abbreviation*	761	59%	0	0%
Other variations				
J *Partially incorrect*	11	1%	79	12%
T *Plus/Minus*	23	2%	13	2%
X *Stop word*	4	0%	17	3%
complex				
Not assessable				
C *Different language*	232	60%	170	21%
Z *Not available*	0	0%	484	60%
Missing data values				
E *Omitted*	115	30%	90	11%
P *No author name*	0	0%	0	0%
Completely incorrect				
D *Completely incorrect*	41	10%	64	8%
SUM	1,728		1,589	

Table 52: Frequency of IACs – General Medicine

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	3	1%	5	1%
L *Informational letter*	2	1%	2	1%
N *Additional information*	26	11%	61	19%
S *Padded*	12	5%	60	19%
Disarranged data values				
G *Interchanged fields*	9	4%	13	4%
H *Jumbled value*	19	8%	19	6%
O *Incorrect order of authors*	162	70%	162	50%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	23	1%	39	3%
V *Incorrect interpretation of add. information*	5	0%	3	0%
Spelling variations				
A *Typographical variation*	35	2%	46	3%
B *Spelling error*	85	5%	111	7%
Q *Special character*	169	9%	134	9%
Y *Word stem*	2	0%	42	3%
Abbreviated data values				
F *Cropped*	866	45%	904	60%
I *Abbreviation*	633	33%	68	4%
Other variations				
J *Partially incorrect*	17	1%	78	5%
T *Plus/Minus*	63	3%	41	3%
X *Stop word*	11	1%	47	3%
complex				
Not assessable				
C *Different language*	293	45%	189	36%
Missing data values				
E *Omitted*	310	47%	256	48%
Completely incorrect				
D *Completely incorrect*	53	8%	87	16%
SUM	2,798		2,367	

Table 53: Frequency of IACs – Educational Science

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	2	4%	1	3%
L *Informational letter*	14	29%	14	34%
N *Additional information*	3	6%	3	7%
S *Padded*	15	30%	11	27%
Disarranged data values				
G *Interchanged fields*	14	29%	12	29%
H *Jumbled value*	1	2%	0	0%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	3	2%	3	1%
Spelling variations				
A *Typographical variation*	3	2%	0	0%
B *Spelling error*	18	10%	8	3%
Q *Special character*	8	4%	140	48%
Y *Word stem*	22	12%	19	7%
Abbreviated data values				
F *Cropped*	64	35%	46	16%
I *Abbreviation*	10	5%	6	2%
Other variations				
J *Partially incorrect*	9	5%	21	7%
T *Plus/Minus*	40	22%	39	14%
X *Stop word*	7	3%	7	2%
complex				
Not assessable				
C *Different language*	5	6%	109	45%
Missing data values				
E *Omitted*	66	77%	67	28%
Completely incorrect				
D *Completely incorrect*	15	17%	66	27%
SUM	319		572	

Table 54: Frequency of IACs – Political Science

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	3	7%	3	7%
L *Informational letter*	16	35%	16	36%
N *Additional information*	3	7%	3	7%
S *Padded*	6	13%	4	9%
Disarranged data values				
G *Interchanged fields*	8	17%	8	18%
H *Jumbled value*	2	4%	2	5%
O *Incorrect order of authors*	8	17%	8	18%
moderate				
Spelling variations				
A *Typographical variation*	1	0%	0	0%
B *Spelling error*	30	11%	14	5%
Q *Special character*	10	4%	89	31%
Y *Word stem*	7	3%	5	2%
Abbreviated data values				
F *Cropped*	168	60%	130	45%
I *Abbreviation*	26	9%	20	7%
Other variations				
J *Partially incorrect*	6	2%	6	2%
T *Plus/Minus*	21	8%	21	7%
X *Stop word*	8	3%	4	1%
complex				
Not assessable				
C *Different language*	0	0%	181	41%
Z *Not available*	0	0%	100	23%
Missing data values				
E *Omitted*	94	59%	92	21%
P *No author name*	7	4%	7	2%
Completely incorrect				
D *Completely incorrect*	59	37%	56	13%
SUM	483		769	

Table 55: Frequency of IACs – Sociology

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	3	8%	1	2%
L *Informational letter*	14	37%	14	31%
S *Padded*	5	13%	5	11%
Disarranged data values				
G *Interchanged fields*	15	39%	25	54%
H *Jumbled value*	1	3%	1	2%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	2	1%	2	0%
Spelling variations				
A *Typographical variation*	7	2%	1	0%
B *Spelling error*	33	10%	34	11%
Q *Special character*	2	1%	130	41%
Y *Word stem*	3	1%	2	0%
Abbreviated data values				
F *Cropped*	140	43%	119	38%
I *Abbreviation*	111	34%	6	2%
Other variations				
J *Partially incorrect*	9	3%	8	3%
T *Plus/Minus*	11	3%	5	2%
X *Stop word*	8	2%	9	3%
complex				
Not assessable				
C *Different language*	6	3%	110	36%
Z *Not available*	84	42%	85	28%
Missing data values				
E *Omitted*	74	37%	68	22%
P *No author name*	11	6%	11	3%
Completely incorrect				
D *Completely incorrect*	23	12%	35	11%
SUM	562		671	

Table 56: Overall descriptive statistics – Language of cited article

	Orig-Ref		WoS-Ref	
	Eng	Ger	Eng	Ger
No of citing references	2,234	1,695	2,234	1,695
No of assessed data values	30,417	24,411	30,471	24,390
No of inaccuracies	4,105	4,003	3,761	4,414

Table 57: Frequency of IACs – English cited articles

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	13	5%	12	3%
L *Informational letter*	24	9%	22	6%
N *Additional information*	5	2%	6	2%
S *Padded*	23	9%	28	7%
Disarranged data values				
G *Interchanged fields*	56	21%	57	15%
H *Jumbled value*	5	2%	6	2%
O *Incorrect order of authors*	137	52%	245	65%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	72	3%	139	9%
V *Incorrect interpretation of add. information*	3	0%	3	0%
Spelling variations				
A *Typographical variation*	35	1%	34	2%
B *Spelling error*	114	5%	128	8%
Q *Special character*	50	2%	101	6%
Y *Word stem*	27	1%	27	2%
Abbreviated data values				
F *Cropped*	945	37%	827	52%
I *Abbreviation*	1,147	45%	196	12%
Other variations				
J *Partially incorrect*	29	1%	34	2%
T *Plus/Minus*	89	4%	81	5%
X *Stop word*	24	1%	25	2%
complex				
Not assessable				
C *Different language*	0	0%	3	0%
Z *Not available*	0	0%	414	23%
Missing data values				
E *Omitted*	1,131	87%	1,144	64%
P *No author name*	11	1%	11	1%
Completely incorrect				
D *Completely incorrect*	165	12%	218	12%
SUM	4,105		3,761	

Table 58: Frequency of IACs – German cited articles

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	3	1%	3	1%
L *Informational letter*	24	9%	24	4%
N *Additional information*	33	12%	255	47%
S *Padded*	31	11%	73	14%
Disarranged data values				
G *Interchanged fields*	28	10%	38	7%
H *Jumbled value*	26	10%	21	4%
O *Incorrect order of authors*	126	47%	124	23%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	5	0%	5	0%
V *Incorrect interpretation of add. information*	2	0%	0	0%
Spelling variations				
A *Typographical variation*	20	1%	14	1%
B *Spelling error*	151	7%	181	9%
Q *Special character*	250	11%	721	35%
Y *Word stem*	11	0%	57	3%
Abbreviated data values				
F *Cropped*	702	31%	598	29%
I *Abbreviation*	1,023	45%	163	8%
Other variations				
J *Partially incorrect*	25	1%	194	10%
T *Plus/Minus*	77	3%	46	2%
X *Stop word*	17	1%	65	3%
complex				
Not assessable				
C *Different language*	549	38%	784	43%
Z *Not available*	84	6%	255	14%
Missing data values				
E *Omitted*	746	51%	683	37%
P *No author name*	12	1%	12	1%
Completely incorrect				
D *Completely incorrect*	58	4%	98	5%
SUM	4,003		4,414	

Table 59: Overall descriptive statistics – document types

		No. of cited ref.	No. of data values	No of inacc.
Assessment result Orig-Ref	Article	3,039	42,075	6,139
	Review	479	7,137	1,101
	Proc.Pap	198	2,712	446
	EditMat	104	1,416	169
	Letter	54	882	151
	Book/Ch	51	534	90
	Other	4	72	12
Assessment result WoS-Ref	Article	3,039	42,099	6,275
	Review	479	7,146	1,032
	Proc.Pap	198	2,712	382
	EditMat	104	1,416	189
	Letter	54	882	155
	Book/Ch	51	534	133
	Other	4	72	9

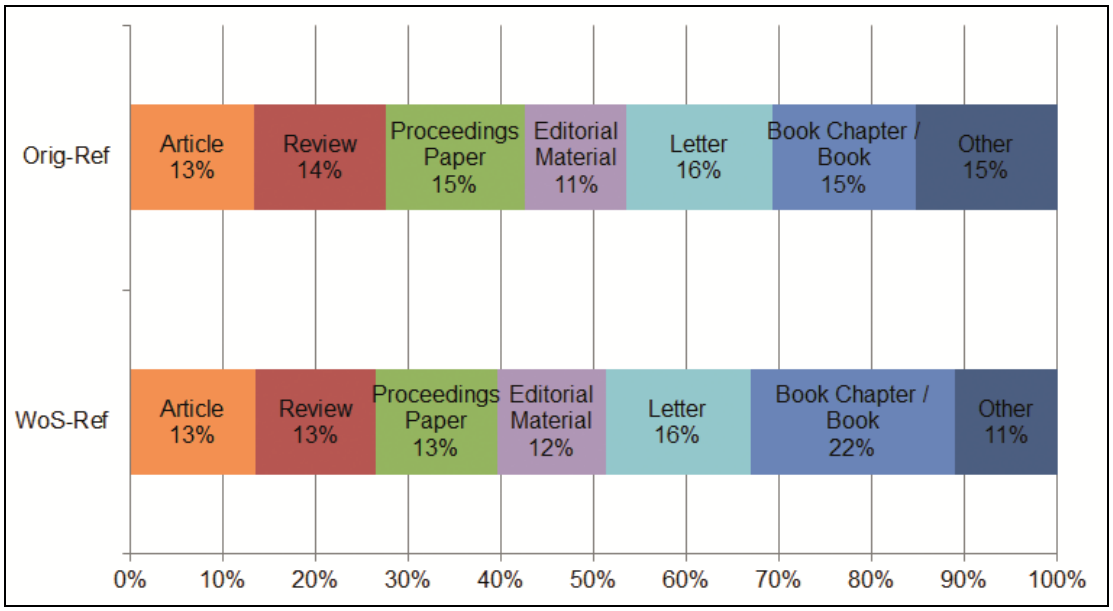


Figure 33: Shares of inaccuracies per document type

Table 60: Frequency of IACs – Article

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	15	4%	14	2%
L *Informational letter*	34	8%	33	5%
N *Additional information*	25	6%	182	26%
S *Padded*	47	11%	92	13%
Disarranged data values				
G *Interchanged fields*	71	17%	80	11%
H *Jumbled value*	26	6%	23	3%
O *Incorrect order of authors*	204	48%	286	40%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	66	2%	117	4%
V *Incorrect interpretation of add. information*	5	0%	3	0%
Spelling variations				
A *Typographical variation*	38	1%	25	1%
B *Spelling error*	199	6%	242	9%
Q *Special character*	220	6%	670	25%
Y *Word stem*	32	1%	71	3%
Abbreviated data values				
F *Cropped*	1,133	32%	973	36%
I *Abbreviation*	1,639	46%	270	10%
Other variations				
J *Partially incorrect*	41	1%	167	6%
T *Plus/Minus*	139	4%	100	4%
X *Stop word*	30	1%	66	2%
complex				
Not assessable				
C *Different language*	413	19%	604	21%
Z *Not available*	70	3%	506	18%
Missing data values				
E *Omitted*	1,520	70%	1,499	53%
P *No author name*	11	1%	11	0%
Completely incorrect				
D *Completely incorrect*	161	7%	241	8%
SUM	6,139		6,275	

Table 61: Frequency of IACs – Review

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
L *Informational letter*	6	11%	5	4%
N *Additional information*	11	20%	54	44%
S *Padded*	2	4%	5	4%
Disarranged data values				
G *Interchanged fields*	8	15%	8	6%
H *Jumbled value*	1	2%	1	1%
O *Incorrect order of authors*	26	48%	50	41%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	9	1%	21	4%
Spelling variations				
A *Typographical variation*	8	1%	17	3%
B *Spelling error*	32	4%	33	6%
Q *Special character*	48	7%	65	12%
Y *Word stem*	1	0%	2	0%
Abbreviated data values				
F *Cropped*	268	38%	252	48%
I *Abbreviation*	316	45%	72	14%
Other variations				
J *Partially incorrect*	4	1%	40	8%
T *Plus/Minus*	13	2%	14	3%
X *Stop word*	5	1%	11	2%
complex				
Not assessable				
C *Different language*	86	25%	67	18%
Z *Not available*	4	1%	62	16%
Missing data values				
E *Omitted*	228	66%	219	57%
P *No author name*	2	1%	2	1%
Completely incorrect				
D *Completely incorrect*	23	7%	32	8%
SUM	1,101		1,032	

Table 62: Frequency of IACs – Proceedings paper

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	1	6%	1	3%
L *Informational letter*	6	35%	6	19%
N *Additional information*	1	6%	13	42%
S *Padded*	2	12%	4	13%
Disarranged data values				
G *Interchanged fields*	1	6%	2	7%
H *Jumbled value*	2	12%	1	3%
O *Incorrect order of authors*	4	23%	4	13%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	2	0%	4	2%
V *Incorrect interpretation of add. information*	0	0%	0	0%
Spelling variations				
A *Typographical variation*	5	2%	4	2%
B *Spelling error*	20	7%	22	11%
Q *Special character*	24	8%	26	13%
Y *Word stem*	1	0%	3	2%
Abbreviated data values				
F *Cropped*	127	42%	107	53%
I *Abbreviation*	115	38%	10	5%
Other variations				
J *Partially incorrect*	3	1%	13	6%
T *Plus/Minus*	6	2%	7	3%
X *Stop word*	2	0%	6	3%
complex				
Not assessable				
C *Different language*	29	24%	30	20%
Z *Not available*	4	3%	39	26%
Missing data values				
E *Omitted*	76	61%	64	43%
Completely incorrect				
D *Completely incorrect*	15	12%	16	11%
SUM	446		382	

Table 63: Frequency of IACs – Editorial material

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
N *Additional information*	0	0%	5	38%
S *Padded*	0	0%	0	0%
Disarranged data values				
G *Interchanged fields*	1	14%	1	8%
H *Jumbled value*	0	0%	1	8%
O *Incorrect order of authors*	6	86%	6	46%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	0	0%	2	2%
Spelling variations				
A *Typographical variation*	1	1%	1	1%
B *Spelling error*	5	4%	7	8%
Q *Special character*	5	4%	27	30%
Y *Word stem*	2	2%	5	6%
Abbreviated data values				
F *Cropped*	48	40%	36	40%
I *Abbreviation*	47	40%	0	0%
Other variations				
J *Partially incorrect*	3	3%	3	3%
T *Plus/Minus*	5	4%	4	5%
X *Stop word*	2	2%	4	5%
complex				
Not assessable				
C *Different language*	9	20%	31	36%
Z *Not available*	1	2%	22	25%
Missing data values				
E *Omitted*	28	64%	24	28%
Completely incorrect				
D *Completely incorrect*	6	14%	10	11%
SUM	169		189	

Table 64: Frequency of IACs – Letter

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
N *Additional information*	1	4%	5	17%
S *Padded*	3	10%	0	0%
Disarranged data values				
H *Jumbled value*	1	4%	1	3%
O *Incorrect order of authors*	23	82%	23	80%
moderate				
Spelling variations				
A *Typographical variation*	1	1%	1	1%
B *Spelling error*	5	6%	4	5%
Q *Special character*	2	2%	15	20%
Y *Word stem*	1	1%	2	3%
Abbreviated data values				
F *Cropped*	42	45%	44	59%
I *Abbreviation*	37	40%	5	7%
Other variations				
J *Partially incorrect*	3	3%	3	4%
T *Plus/Minus*	2	2%	0	0%
X *Stop word*	0	0%	1	1%
complex				
Not assessable				
C *Different language*	9	30%	15	29%
Z *Not available*	0	0%	17	33%
Missing data values				
E *Omitted*	7	23%	5	10%
Completely incorrect				
D *Completely incorrect*	14	47%	14	28%
SUM	151		155	

Table 65: Frequency of IACs – Book / Book Chapter

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
L *Informational letter*	2	33%	2	25%
N *Additional information*	0	0%	2	25%
Disarranged data values				
G *Interchanged fields*	3	50%	4	50%
H *Jumbled value*	1	17%	0	0%
moderate				
Spelling variations				
A *Typographical variation*	2	4%	0	0%
B *Spelling error*	4	9%	1	3%
Q *Special character*	1	2%	19	58%
Y *Word stem*	1	2%	1	3%
Abbreviated data values				
F *Cropped*	22	50%	5	15%
I *Abbreviation*	12	27%	2	6%
Other variations				
J *Partially incorrect*	0	0%	1	3%
T *Plus/Minus*	1	2%	2	6%
X *Stop word*	2	4%	2	6%
complex				
Not assessable				
C *Different language*	2	5%	40	44%
Z *Not available*	5	13%	23	25%
Missing data values				
E *Omitted*	18	46%	16	17%
P *No author name*	10	26%	10	11%
Completely incorrect				
D *Completely incorrect*	4	10%	3	3%
SUM	90		133	

Table 66: Frequency of IACs – Other document types

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
moderate				
Abbreviated data values				
F *Cropped*	7	64%	8	89%
I *Abbreviation*	4	36%	0	0%
Other variations				
J *Partially incorrect*	0	0%	1	11%
complex				
Not assessable				
C *Different language*	1	100%	0	0%
SUM	12		9	

Table 67: Overall descriptive statistics – Language of citing article

		No. of cited ref.	No. of data values	No of inacc.
Assessment result Orig-Ref	English	3,232	45,240	6,819
	German	629	8,592	1,094
	French	23	330	62
	Spanish	16	219	32
	Other	29	447	101
Assessment result WoS-Ref	English	3,232	45,282	6,344
	German	629	8,583	1,661
	French	23	330	47
	Spanish	16	219	31
	Other	29	447	92

Table 68: Distribution of citing articles per language

Citing Article Language	No of references
Chinese	8
Croatian	1
Czech	2
Dutch	2
English	3,232
English, Spanish	1
French	23
German	629
Italian	4
Japanese	1
Korean	1
Lithuanian	1
Portuguese	2
Serbian	1
Spanish	16
Turkish	5

Table 69: Frequency of IACs – English citing articles

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	14	4%	13	2%
L *Informational letter*	30	8%	28	4%
N *Additional information*	36	10%	251	34%
S *Padded*	28	8%	76	10%
Disarranged data values				
G *Interchanged fields*	68	19%	73	10%
H *Jumbled value*	15	4%	12	2%
O *Incorrect order of authors*	171	47%	278	38%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	73	2%	138	5%
V *Incorrect interpretation of add. information*	2	0%	2	0%
Spelling variations				
A *Typographical variation*	41	1%	46	2%
B *Spelling error*	226	5%	282	10%
Q *Special character*	245	6%	382	13%
Y *Word stem*	33	1%	75	3%
Abbreviated data values				
F *Cropped*	1,409	35%	1,248	43%
I *Abbreviation*	1,813	45%	326	11%
Other variations				
J *Partially incorrect*	41	1%	215	7%
T *Plus/Minus*	127	3%	98	3%
X *Stop word*	30	1%	80	3%
complex				
Not assessable				
C *Different language*	475	20%	277	10%
Z *Not available*	30	1%	491	18%
Missing data values				
E *Omitted*	1,710	71%	1,673	62%
P *No author name*	13	0%	13	0%
Completely incorrect				
D *Completely incorrect*	189	8%	267	10%
SUM	6,819		6,344	

Table 70: Frequency of IACs – German citing articles

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
L *Informational letter*	17	10%	17	10%
N *Additional information*	2	1%	7	4%
S *Padded*	23	14%	18	11%
Disarranged data values				
G *Interchanged fields*	15	9%	21	12%
H *Jumbled value*	16	10%	15	9%
O *Incorrect order of authors*	92	56%	91	54%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	2	0%	2	0%
V *Incorrect interpretation of add. information*	3	0%	1	0%
Spelling variations				
A *Typographical variation*	14	2%	2	0%
B *Spelling error*	35	5%	20	3%
Q *Special character*	52	8%	437	65%
Y *Word stem*	3	0%	6	1%
Abbreviated data values				
F *Cropped*	194	29%	142	21%
I *Abbreviation*	313	47%	26	4%
Other variations				
J *Partially incorrect*	13	2%	9	1%
T *Plus/Minus*	38	6%	28	4%
X *Stop word*	8	1%	7	1%
complex				
Not assessable				
C *Different language*	61	24%	509	63%
Z *Not available*	54	21%	158	19%
Missing data values				
E *Omitted*	102	40%	91	11%
P *No author name*	7	3%	7	1%
Completely incorrect				
D *Completely incorrect*	30	12%	47	6%
SUM	1,094		1,661	

Table 71: Frequency of IACs – French citing articles

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	1	100%	1	33%[52]
N *Additional information*	0	0%	1	33%
S *Padded*	0	0%	1	33%
moderate				
Spelling variations				
A *Typographical variation*	0	0%	0	0%
B *Spelling error*	2	5%	3	13%
Q *Special character*	1	2%	2	9%
Abbreviated data values				
F *Cropped*	22	50%	13	57%
I *Abbreviation*	17	39%	0	0%
Other variations				
J *Partially incorrect*	0	0%	3	13%
T *Plus/Minus*	1	2%	1	4%
X *Stop word*	1	2%	1	4%
complex				
Not assessable				
C *Different language*	5	29%	1	5%
Z *Not available*	0	0%	10	48%
Missing data values				
E *Omitted*	10	59%	8	38%
Completely incorrect				
D *Completely incorrect*	2	12%	2	9%
SUM	62		47	

[52] The numbers in this subcategory add up to 99% due the rounding. The exact shares are 33.33333% each.

Table 72: Frequency of IACs – Spanish citing articles

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
L *Informational letter*	1	50%	1	25%
S *Padded*	0	0%	2	50%
Disarranged data values				
G *Interchanged fields*	1	50%	1	25%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	2	10%	4	19%
Spelling variations				
Q *Special character*	0	0%	1	5%
Y *Word stem*	1	4%	1	5%
Abbreviated data values				
F *Cropped*	8	38%	10	48%
I *Abbreviation*	8	38%	2	9%
Other variations				
J *Partially incorrect*	0	0%	1	5%
X *Stop word*	2	10%	2	9%
complex				
Not assessable				
C *Different language*	4	45%	0	0%
Z *Not available*	0	0%	1	17%
Missing data values				
E *Omitted*	2	22%	2	33%
P *No author name*	3	33%	3	50%
SUM	32		31	

Table 73: Frequency of IACs – Citing articles in Other languages

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	1	25%	1	14%
N *Additional information*	0	0%	2	29%
S *Padded*	3	75%	4	57%
moderate				
Spelling variations				
B *Spelling error*	2	5%	4	17%
Q *Special character*	2	5%	0	0%
Y *Word stem*	1	3%	2	9%
Abbreviated data values				
F *Cropped*	14	37%	12	52%
I *Abbreviation*	19	50%	5	22%
complex				
Not assessable				
C *Different language*	4	7%	0	0%
Z *Not available*	0	0%	9	15%
Missing data values				
E *Omitted*	53	90%	53	85%
Completely incorrect				
D *Completely incorrect*	2	3%	0	0%
SUM	101		92	

Table 74: Overall descriptive statistics – Citation windows

	Assessment result Orig-Ref			Assessment result WoS-Ref		
	1998-2002	2003-2007	2008-2012	1998-2002	2003-2007	2008-2012
No of citing references	629	1,384	1,916	629	1,384	1,916
No of data values	9,102	20,208	25,518	9,141	20,211	25,509
No of inaccuracies	1,600	3,118	3,390	1,670	3,015	3,490

Table 75: Frequency of IACs – Citation window 1998-2002

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	1	1%	1	0%
L *Informational letter*	9	6%	9	3%
N *Additional information*	7	5%	77	27%
S *Padded*	14	10%	7	3%
Disarranged data values				
G *Interchanged fields*	9	6%	9	3%
O *Incorrect order of authors*	103	72%	180	64%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	18	2%	40	6%
Spelling variations				
A *Typographical variation*	6	1%	3	0%
B *Spelling error*	65	7%	77	12%
Q *Special character*	51	6%	118	19%
Y *Word stem*	6	1%	17	3%
Abbreviated data values				
F *Cropped*	241	28%	204	32%
I *Abbreviation*	438	51%	103	16%
Other variations				
J *Partially incorrect*	7	1%	39	6%
T *Plus/Minus*	17	2%	15	3%
X *Stop word*	6	1%	20	3%
complex				
Not assessable				
C *Different language*	71	12%	145	19%
Z *Not available*	13	2%	78	11%
Missing data values				
E *Omitted*	479	80%	472	63%
P *No author name*	2	0%	2	0%
Completely incorrect				
D *Completely incorrect*	37	6%	54	7%
SUM	1,600		1,670	

Table 76: Frequency of IACs – Citation window 2003-2007

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	8	4%	7	2%
L *Informational letter*	7	4%	7	2%
N *Additional information*	14	8%	92	31%
S *Padded*	19	10%	39	13%
Disarranged data values				
G *Interchanged fields*	36	19%	38	13%
H *Jumbled value*	15	8%	13	4%
O *Incorrect order of authors*	86	47%	104	35%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	31	2%	57	4%
V *Incorrect interpretation of add. information*	2	0%	1	0%
Spelling variations				
A *Typographical variation*	23	1%	23	2%
B *Spelling error*	104	5%	130	9%
Q *Special character*	125	7%	305	22%
Y *Word stem*	9	0%	25	2%
Abbreviated data values				
F *Cropped*	670	35%	584	41%
I *Abbreviation*	844	45%	130	9%
Other variations				
J *Partially incorrect*	26	1%	80	6%
T *Plus/Minus*	55	3%	46	3%
X *Stop word*	10	1%	27	2%
complex				
Not assessable				
C *Different language*	205	20%	267	20%
Z *Not available*	27	3%	226	17%
Missing data values				
E *Omitted*	723	70%	699	54%
P *No author name*	7	0%	7	1%
Completely incorrect				
D *Completely incorrect*	72	7%	108	8%
SUM	3,118		3,015	

Table 77: Frequency of IACs – Citation window 2008-2012

	Orig-Ref		WoS-Ref	
IAC	Count	Type %	Count	Type %
simple				
Added data values				
K *Space*	7	3%	7	2%
L *Informational letter*	32	16%	30	9%
N *Additional information*	17	8%	92	28%
S *Padded*	21	10%	55	17%
Disarranged data values				
G *Interchanged fields*	39	19%	48	14%
H *Jumbled value*	16	8%	14	4%
O *Incorrect order of authors*	74	36%	85	26%
moderate				
Incorrect interpretation of data values				
M *Incorrect interpretation of author names*	28	1%	47	3%
V *Incorrect interpretation of add. information*	3	0%	2	0%
Spelling variations				
A *Typographical variation*	26	1%	22	1%
B *Spelling error*	96	5%	102	6%
Q *Special character*	124	6%	399	25%
Y *Word stem*	23	1%	42	3%
Abbreviated data values				
F *Cropped*	736	36%	637	40%
I *Abbreviation*	888	43%	126	8%
Other variations				
J *Partially incorrect*	21	1%	109	7%
T *Plus/Minus*	94	5%	66	4%
X *Stop word*	25	1%	43	3%
complex				
Not assessable				
C *Different language*	273	25%	375	24%
Z *Not available*	44	4%	365	23%
Missing data values				
E *Omitted*	675	60%	656	42%
P *No author name*	14	1%	14	1%
Completely incorrect				
D *Completely incorrect*	114	10%	154	10%
SUM	3,390		3,490	

G False positive matches in WoS

Appendix G summarizes the false positive matches identified in WoS.

WoS-UT	WoS category	ID cited article	comment
000270020200021	Business & Economics; Computer Science; Public Administration; Telecommunications	BeSo08_008	
000286873600006	Family Studies; Government & Law; Social Work	BeSo98_001	
000181755500018	Neurosciences & Neurology; Pharmacology & Pharmacy; Psychiatry	BeSo98_001	different domain
000084485900003	Biochemistry & Molecular Biology; Cell Biology	BeSo98_001	different domain
000228255900003	Dentistry, Oral Surgery & Medicine	BeSo98_002	different domain
000174834200012	Biochemistry & Molecular Biology; Endocrinology & Metabolism; Toxicology; Zoology	BeSo98_002	different domain
000167374800009	Government & Law	BeSo98_004	
000090094400001	Government & Law	BeSo98_004	
000087564800012	Government & Law	BeSo98_004	
000084295000021	Life Sciences & Biomedicine - Other Topics	BeSo98_010	different domain
000083892100025	Chemistry	HAC98_001	erratum
000222356900009	Anthropology	HaCl98_009	different domain
000224959600001	Zoology	HaCl98_009	
000222091700004	Evolutionary Biology; Zoology	HaCl98_009	
000187712000001	Entomology	HaCl98_009	
000178148400003	Entomology	HaCl98_009	
000171172900018	Entomology	HaCl98_009	
000168258800013	Entomology	HaCl98_009	
000086539300009	Entomology	HaCl98_009	
000082711000013	Physics	HaCl98_009	
000241359000097	Computer Science	JCuSt98_005	
000240091500045	Computer Science	JCuSt98_005	
000185510800036	Computer Science	JCuSt98_005	

WoS-UT	WoS category	ID cited article	comment
000171566400006	Biochemistry & Molecular Biology; Biotechnology & Applied Microbiology; Computer Science; Mathematical & Computational Biology; Mathematics	JCuSt98_005	different domain
000185853600015	Internal & General Medicine	JTM03_005	erratum
000263372000008	Computer Science	JTM03_010	
000262588100046	Oncology; Obstetrics & Gynecology	PoTh08_002	different domain
000243443300005	Oncology; Genetics & Heredity	PoTh98_007	different domain
000234871900003	Ethics; Psychology, Multidisciplinary	PoVi03_006	
000240470700011	General & Internal Medicine	WDMW98_001	
000181820100013	Endocrinology & Metabolism	WDMW98_005	
000180534800006	Cardiovascular System & Cardiology	WDMW98_005	
000181024300001	Orthopedics	WOrth98_001	

H Irregular WoS records

Appendix H summarizes all irregular records identified in WoS.

Table 78: WoS target article with an incorrect article language

Internal ID	WoS-UT	WoS Article language	Correct Article language
BeSo98_005	000076443500008	English	German
PoVi98_010	000076743200005	French	German

Table 79: WoS source articles with an incorrect article language

Internal ID	WoS-UT	WoS Article language	Correct Article language
BeSo03_016	000250299500003	English	German
HAC03_038	000243493400002	Russian	English
HAC03_170	000237964900004	Russian	English
HaCl03_165	000288842900008	Spanish	English
JCuSt08_047	000295191100011	English	Spanish
JTM03_106	000285366400005	English	English; Spanish
JTM98_131	000225649800008	English	French
PoVi03_059	000250299500002	English	German
PoVi03_081	000314511000003	English	German
SoIn98_030	000261211400002	Hungarian	English
WCZ98_225	000229379500004	Russian	English
WCZ98_277	000223168100001	German	English
WDMW03_039	000254395800002	English	German
WDMW03_189	000236678200008	French	English
WDMW98_057	000183950500008	English	German
WDMW98_158	000079710500013	English	German
WDMW98_221	000221717600005	English; Estonian	German
WDMW98_285	000249522200024	Spanish	English
WOrth03_166	000229048400007	English	German
ZPad98_029	000248753700003	English	German

Table 80: WoS target articles with missing ending page numbers

Internal ID	WoS-UT	WoS EP	Correct EP
BeSo03_001	000185021400005	+	238
BeSo03_002	000189260500005	+	529
BeSo03_003	000189260500008	+	584
BeSo03_004	000186419500002	+	323
BeSo03_005	000189260500004	+	510
BeSo03_006	000186419500005	+	393
BeSo03_007	000189260500003	+	495
BeSo03_008	000182397500005	+	96
BeSo03_009	000185021400003	+	195
BeSo03_010	000185021400007	+	274
BeSo98_001	000073119800002	+	22
BeSo98_002	000076443500007	+	380
BeSo98_003	000077822800007	+	547
BeSo98_004	000076443500010	+	420
BeSo98_005	000076443500008	+	392
BeSo98_006	000073119800009	+	142
BeSo98_007	000076443500006	+	357
BeSo98_008	000077822800004	+	505
BeSo98_009	000074825600002	+	180
BeSo98_010	000074825600004	+	222
HaCl03_001	000184794300022	+	538
HaCl03_002	000184794300003	+	369
HaCl03_003	000183799700006	+	239
HaCl03_004	000181956900005	+	49
HaCl03_005	000181956900007	+	71
HaCl03_006	000181956900013	+	175
HaCl03_007	000184794300015	+	470
HaCl03_008	000183799700014	+	316
HaCl03_009	000183799700013	+	308
HaCl03_010	000181956900002	+	15
HaCl98_001	000075899800006	+	383
HaCl98_002	000073670600004	+	176
HaCl98_003	000077576200012	+	645
HaCl98_004	000075899800005	+	370
HaCl98_005	000075899800011	+	450
HaCl98_006	000077576200007	+	578
HaCl98_007	000075899800010	+	429
HaCl98_008	000073670600011	+	277
HaCl98_009	000072590900002	+	15
HaCl98_010	000077576200008	+	587
PoVi03_001	000188869300004	+	528
PoVi03_002	000186430200005	+	369

Internal ID	WoS-UT	WoS EP	Correct EP
PoVi03_003	000182704300005	+	65
PoVi03_004	000188869300002	+	482
PoVi03_006	000182704300004	+	40
PoVi03_007	000184873600002	+	173
PoVi03_008	000186430200003	+	324
PoVi03_009	000184873600003	+	195
PoVi03_010	000186430200006	+	394
PoVi98_001	000076743200004	+	589
PoVi98_002	000078375300001	+	757
PoVi98_005	000078375300002	+	774
PoVi98_007	000075085200002	+	281
PoVi98_008	000073390000004	+	79
PoVi98_009	000073390000003	+	54
WOrth03_001	000184209700004	+	476
WOrth03_005	000183374000012	+	431
WOrth03_006	000184209700010	+	526

Table 81: WoS target articles with a transposed ending page number

Internal ID	WoS-UT	WoS EP	Correct EP
WCZ98_010	000077789300002	300	301
WDMW03_001	000187220400003	2637	2638
WOrth03_008	000184209700014	569	570

Table 82: WoS target articles with incorrect article title

Internal ID	WoS-UT	WoS article title	Correct article title
JCuSt03_003	000180284900002	**OP-ED** Scientific literacy as an emergent feature of collective human praxis	Scientific literacy as an emergent feature of collective human praxis
PoTh03_010	000181590500004	Metaphysics in the dark. A response to Richard Rorty and Ernesto La**d**au	Metaphysics in the dark. A Response to Richard Rorty and Ernesto La**cl**au
ZPad08_004	000261119900002	Test-Based Educational Accountability	Test-Based Educational Accountability. **Research Evidence and Implications**

Table 83: WoS target articles with incorrect or discrepant author names

Internal ID	WoS-UT	WoS		correct	
		author no.	LN	author no.	Author name
HAC03_009	000181850000008	1	Baul, TSB	1	Basu Baul, TS
HAC98_002	000073141300009	4	Malar, EJP	4	Padma Malar, EJ
HAC98_007	000076154100012	1	Hoa, N	1	Tran Huy, NH
HAC98_007	000076154100012	2	Huy, T	2	Ricard, L
HAC98_007	000076154100012	3	Ricard, L	3	Mathey, F
HAC98_007	000076154100012	4	Mathey, F	4	-
HaCl03_002	000184794300003	7	Meyer, PTA	7	Meyer, A
HaCl98_004	000075899800005	2	Viegas, SE	4	Viegas, SF
JCuSt03_002	000185914500001	3	Spillane, JP[53]	3	Jita, L
JTM03_003	000183270000005	12	Knobloch, SJ	12	Knobloch, S
JTM03_008	000181249600002	8	Fradet, MD	8	Douville Fradet, M
JTM03_009	000181677600004	2	Saillour, MF	2	Flament Saillour, M
SoIn98_006	000078438900003	2	Austin, SF	2	-
WOrth98_004	000072512500006	2	Linhard, W	2	Linhart, W
ZPad03_003	000183300700002	3	Feinstein, S	3	Feinstein, L

[53] At the time of downloading the records, this author name was given as the third. When we checked the records again in August, 2014, the third author had been removed but not replaced with the correct third author.

I MISSED CITATIONS

Appendix I documents all information regarding missed citations.

Table 84: Four citations missed by all six data sources

ID	Cited reference information – missed citation	Cited reference information – matched citation
HAC98_214	SHANMUGASUNDARA.M, 1998, HETEROATOM CHEM, P327	Raghunathan R, 1998, HETEROATOM CHEM, V9, P327, DOI 10.1002/(SICI)1098-1071(1998)9:3<327::AID-HC9>3.0.CO;2-6
PoTh08_108	ROOVER J, 2008, POLITICAL THEORY	De Roover J, 2008, POLIT THEORY, V36, P523, DOI 10.1177/0090591708317969
WCZ98_286	ARDUENGO AJ, IN PRESS CHEM UNSERE	Arduengo AJ, 1998, CHEM UNSERER ZEIT, V32, P6, DOI 10.1002/ciuz.19980320103
ZPad08_050	PANT HA, Z PADAGOGIK IN PRESS	Pant HA, 2008, Z PADAGOGIK, V54, P827

Table 85: Overall descriptive statistics – inaccuracies in missed citations

	Orig-Ref	WoS-Ref	CitedRef-WoS	GS	CitedRef-Sco	CWTS	iFQ	SM
No of missed citations	219	219	219	79	58	51	45	199
No of inaccuracies	346	342	365	138	93	108	104	332

Table 86: Overall frequency of IACs in missed citations – Orig-Ref

Assessment result Orig-Ref		
IAC	Count	Type %
simple		
Added data values		
L *Informational letter*	4	8%
N *Additional information*	3	6%
S *Padded*	3	6%
Disarranged data values		
G *Interchanged fields*	30	62%
H *Jumbled value*	5	10%
O *Incorrect order of authors*	4	8%
moderate		
Incorrect interpretation of data values		
M *Incorrect interpretation of author names*	2	1%
Spelling variations		
B *Spelling error*	16	10%
Q *Special character*	6	4%
Abbreviated data values		
F *Cropped*	4	2%
I *Abbreviation*	90	54%
Other variations		
T *Plus/Minus*	49	29%
complex		
Missing data values		
E *Omitted*	52	40%
P *No author name*	18	14%
Completely incorrect		
D *Completely incorrect*	60	46%
SUM	346	

Table 87: Overall frequency of IACs in missed citations – WoS-Ref

Assessment result WoS-Ref		
IAC	Count	Type %
simple		
Added data values		
L *Informational letter*	3	4%
N *Additional information*	11	17%
S *Padded*	3	4%
Disarranged data values		
G *Interchanged fields*	40	61%
H *Jumbled value*	5	8%
O *Incorrect order of authors*	4	6%
moderate		
Incorrect interpretation of data values		
M *Incorrect interpretation of author names*	2	1%
Spelling variations		
B *Spelling error*	17	12%
Q *Special character*	58	40%
Abbreviated data values		
F *Cropped*	4	3%
I *Abbreviation*	11	8%
Other variations		
T *Plus/Minus*	51	36%
complex		
Missing data values		
E *Omitted*	52	39%
P *No author name*	18	14%
Completely incorrect		
D *Completely incorrect*	63	47%
SUM	342	

Table 88: Overall frequency of IACs in missed citations – CitedRef-WoS

Assessment result CitedRef-WoS		
IAC	Count	Type %
simple		
Added data values		
K *Space*	3	3%
N *Additional information*	11	9%
R *Punctuation*	37	32%
S *Padded*	3	3%
U *Full first name*	11	9%
Disarranged data values		
G *Interchanged fields*	45	38%
H *Jumbled value*	4	3%
O *Incorrect order of authors*	4	3%
moderate		
Incorrect interpretation of data values		
M *Incorrect interpretation of author names*	13	11%
Spelling variations		
B *Spelling error*	18	15%
Q *Special character*	5	4%
Abbreviated data values		
F *Cropped*	5	4%
I *Abbreviation*	24	21%
Other variations		
J *Partially incorrect*	1	1%
T *Plus/Minus*	52	44%
complex		
Missing data values		
E *Omitted*	60	46%
P *No author name*	2	2%
Completely incorrect		
D *Completely incorrect*	67	52%
SUM	365	

Table 89: Overall frequency of IACs in missed citations – CitedRef-Sco

Assessment result CitedRef-Sco		
IAC	Count	Type %
simple		
Added data values		
N *Additional information*	7	23%
S *Padded*	1	3%
U *Full first name*	2	6%
Disarranged data values		
G *Interchanged fields*	16	52%
H *Jumbled value*	1	3%
O *Incorrect order of authors*	4	13%
moderate		
Incorrect interpretation of data values		
M *Incorrect interpretation of author names*	7	28%
Spelling variations		
B *Spelling error*	2	8%
Q *Special character*	3	12%
Abbreviated data values		
F *Cropped*	3	12%
Other variations		
T *Plus/Minus*	10	40%
complex		
Missing data values		
E *Omitted*	20	54%
P *No author name*	2	5%
Completely incorrect		
D *Completely incorrect*	15	41%
SUM	93	

Table 90: Overall frequency of IACs in missed citations – GS

Assessment result GS		
IAC	Count	Type %
simple		
Added data values		
L *Informational letter*	3	11%
N *Additional information*	4	14%
S *Padded*	1	4%
Disarranged data values		
G *Interchanged fields*	17	60%
H *Jumbled value*	1	4%
O *Incorrect order of authors*	2	7%
moderate		
Spelling variations		
B *Spelling error*	7	14%
Q *Special character*	11	22%
Abbreviated data values		
F *Cropped*	2	4%
I *Abbreviation*	19	38%
Other variations		
J *Partially incorrect*	0	0%
T *Plus/Minus*	11	22%
complex		
Missing data values		
E *Omitted*	32	54%
P *No author name*	14	23%
Completely incorrect		
D *Completely incorrect*	14	23%
SUM	138	

Table 91: Overall frequency of IACs in missed citations – CWTS

Assessment result CWTS		
IAC	Count	Type %
simple		
Added data values		
N *Additional information*	8	21%
R *Punctuation*	6	16%
S *Padded*	1	2%
U *Full first name*	2	5%
Disarranged data values		
G *Interchanged fields*	17	45%
O *Incorrect order of authors*	4	11%
moderate		
Spelling variations		
B *Spelling error*	3	13%
Abbreviated data values		
F *Cropped*	1	5%
I *Abbreviation*	7	32%
Other variations		
J *Partially incorrect*	1	5%
T *Plus/Minus*	10	45%
complex		
Missing data values		
E *Omitted*	27	56%
Completely incorrect		
D *Completely incorrect*	21	44%
SUM	108	

Table 92: Overall frequency of IACs in missed citations – iFQ

Assessment result iFQ		
IAC	Count	Type %
simple		
Added data values		
K *Space*	1	3%
N *Additional information*	6	20%
R *Punctuation*	16	52%
U *Full first name*	2	6%
Disarranged data values		
G *Interchanged fields*	2	6%
O *Incorrect order of authors*	4	13%
moderate		
Spelling variations		
B *Spelling error*	2	5%
Abbreviated data values		
F *Cropped*	1	23%
I *Abbreviation*	9	23%
Other variations		
J *Partially incorrect*	1	2%
T *Plus/Minus*	27	68%
complex		
Missing data values		
E *Omitted*	23	70%
Completely incorrect		
D *Completely incorrect*	10	30%
SUM	104	

Table 93: Overall frequency of IACs in missed citations – SM

Assessment result SM		
IAC	Count	Type %
simple		
Added data values		
K *Space*	3	3%
N *Additional information*	11	10%
R *Punctuation*	33	31%
S *Padded*	2	2%
U *Full first name*	9	8%
Disarranged data values		
G *Interchanged fields*	41	38%
H *Jumbled value*	4	4%
O *Incorrect order of authors*	4	4%
moderate		
Incorrect interpretation of data values		
M *Incorrect interpretation of author names*	12	11%
Spelling variations		
B *Spelling error*	14	13%
Q *Special character*	5	5%
Abbreviated data values		
F *Cropped*	5	5%
I *Abbreviation*	22	20%
Other variations		
J *Partially incorrect*	1	0%
T *Plus/Minus*	51	46%
complex		
Missing data values		
E *Omitted*	52	45%
Completely incorrect		
D *Completely incorrect*	63	55%
SUM	332	

Table 94: Number of references not matched because of a single inaccuracy (CitedRef-WoS result)

IAC	Last name	First initial	PubYear	Vol no	Starting page	Total
simple						
Added data values						
R *Punctuation*	-	1	-	-	-	1
S *Padded*	-	-	-	1	-	1
U *First full name*	-	1	-	-	-	1
Disarranged data values						
G *Interchanged fields (G1)*	-	-	-	15	5	20
H *Jumbled value*	-	-	-	-	2	2
moderate						
Incorrect interpretation of data values						
M *Incorrect interpretation of author names*	-	1	-	-	-	1
Spelling variations						
B *Spelling error*	8	-	-	-	-	8
Q *Special character*	2	-	-	-	-	2
Abbreviated data values						
F *Cropped*	-	-	-	1	1	2
Other variations						
T *Plus/Minus*	-	-	6	2	13	21
complex						
Missing data values						
E *Omitted*	-	1	-	7	-	8
Completely incorrect						
D *Completely incorrect*	-	1	-	5	29	35
SUM						**102**

Table 95 summarizes the cited reference information of 18 references where the original reference was completely accurate and the inaccuracies caused by the data handling. Figure 34 to Figure 51 give screenshots of the references as they were found in the original citing articles.

Table 95: Cited reference information of missed citing articles without inaccuracies in the original reference

ID	Cited reference information of missed citing article	Cited reference information of matched citing article
BeSo98_062	WESTERN B, 1998, BERL J SOZIOL, V**2**, P159	Western B, 1998, BERL J SOZIOL, V8, P159
HAC03_216	**D**ELBRUNO JJ, 2003, HETEROATOM CHEM, V14, P189	BelBruno JJ, 2003, HETEROATOM CHEM, V14, P189, DOI 10.1002/hc.10127
HAC98_213	RA**T**HUNATHAN R, 1998, HETEROATOM CHEM, V9, P327	Raghunathan R, 1998, HETEROATOM CHEM, V9, P327, DOI 10.1002/(SICI)1098-1071(1998)9:3<327::AID-HC9>3.0.CO;2-6
HaCl98_093	**ELIAS MG**, 1998, HAND CLIN, V14, P165	Garcia-Elias M, 1998, HAND CLIN, V14, P165
HaCl98_094	**ELIAS MG**, 1998, HAND CLIN, V14, P165	Garcia-Elias M, 1998, HAND CLIN, V14, P165
HaCl98_095	**ELIAS MG**, 1998, HAND CLIN, V14, P165	Garcia-Elias M, 1998, HAND CLIN, V14, P165
HaCl98_096	**ELIAS MG**, 1998, HAND CLIN, V14, P165	Garcia-Elias M, 1998, HAND CLIN, V14, P165
HaCl98_097	**ELIAS MG**, 1998, HAND CLIN, V14, P165	Garcia-Elias M, 1998, HAND CLIN, V14, P165
HaCl98_141	SZABO RM, 1998, HAND CLIN, V14, **pR9**	Szabo RM, 1998, HAND CLIN, V14, P419
PoTh03_128	SCOTT D, 2003, POLITICAL THEORY FEB, V**3**, P92	Scott D, 2003, POLIT THEORY, V31, P92, DOI 10.1177/0090591702239440
PoTh03_137	NASSTROM S, 2003, POLITICAL THEORY, V31	Nasstrom S, 2003, POLIT THEORY, V31, P808, DOI 10.1177/0090591703252158
PoTh08_029	Nasstrom S**.**, 2003, POLIT THEORY, V31, P808	Nasstrom S, 2003, POLIT THEORY, V31, P808, DOI 10.1177/0090591703252158
PoTh03_139	DEVEAUX M, 2003, POLIT THEORY, V31, P**781**	Deveaux M, 2003, POLIT THEORY, V31, P780, DOI 10.1177/0090591703256685
PoTh03_140	BADER V, 2003, POLIT THEORY, V31, P**269**	Bader V, 2003, POLIT THEORY, V31, P265, DOI 10.1177/0090591702251012
SoIn03_149	FEATHERSTONE R, 2003, SOCIOL INQ, V73, P**480**	Featherstone R, 2003, SOCIOL INQ, V73, P471, DOI 10.1111/1475-682X.00067
SoIn03_150	**LOMSKYFEDER**, 2003, SOCIOL INQ, V73, P114	Lomsky-Feder E, 2003, SOCIOL INQ, V73, P114, DOI 10.1111/1475-682X.00043
WDMW03_192[54]	**HAUCR** H, 2003, DEUT MED WOCHENSCHR, V128, P2632	Hauner H, 2003, DEUT MED WOCHENSCHR, V128, P2632, DOI 10.1055/s-2003-812396
WDMW03_197	WEIDE R, 2003, DEUT MED WOCHENSCHR, V128**5**, P2418	Weide R, 2003, DEUT MED WOCHENSCHR, V128, P2418, DOI 10.1055/s-2003-43590
ZPad08_047	KLIEME E, 2008, Z PADAGOGIK, P222	Klieme E, 2008, Z PADAGOGIK, V54, P222

[54] The copy of this citing article obtained via interlibrary loan was of particular poor quality. Therefore, we did not scan and include it.

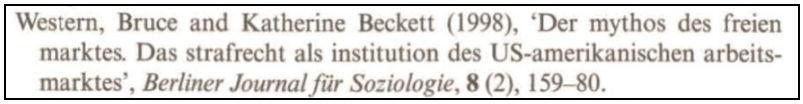

Western, Bruce and Katherine Beckett (1998), 'Der mythos des freien marktes. Das strafrecht als institution des US-amerikanischen arbeits-marktes', *Berliner Journal für Soziologie*, **8** (2), 159–80.

Figure 34: BeSo98_062, citing article

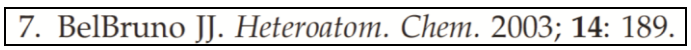

7. BelBruno JJ. *Heteroatom. Chem.* 2003; **14**: 189.

Figure 35: HAC03_216, citing article

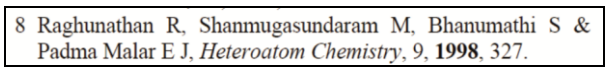

8 Raghunathan R, Shanmugasundaram M, Bhanumathi S & Padma Malar E J, *Heteroatom Chemistry*, 9, **1998**, 327.

Figure 36: HAC98_213, citing article

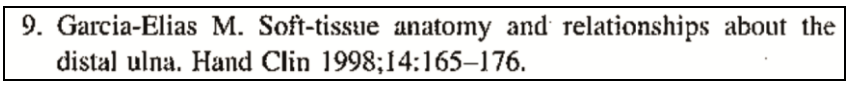

9. Garcia-Elias M. Soft-tissue anatomy and relationships about the distal ulna. Hand Clin 1998;14:165–176.

Figure 37: HaCl98_093, citing article

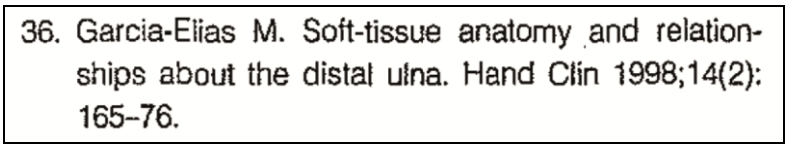

36. Garcia-Elias M. Soft-tissue anatomy and relationships about the distal ulna. Hand Clin 1998;14(2): 165–76.

Figure 38: HaCl98_094, citing article

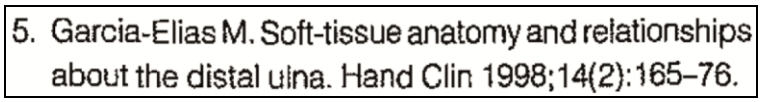

5. Garcia-Elias M. Soft-tissue anatomy and relationships about the distal ulna. Hand Clin 1998;14(2):165–76.

Figure 39: HaCl98_095, citing article

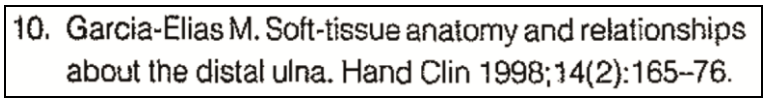

10. Garcia-Elias M. Soft-tissue anatomy and relationships about the distal ulna. Hand Clin 1998;14(2):165–76.

Figure 40: HaCl98_096, citing article

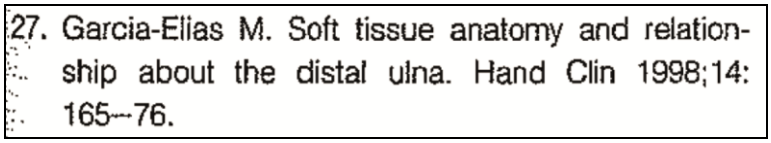

27. Garcia-Elias M. Soft tissue anatomy and relationship about the distal ulna. Hand Clin 1998;14: 165–76.

Figure 41: HaCl98_097, citing article

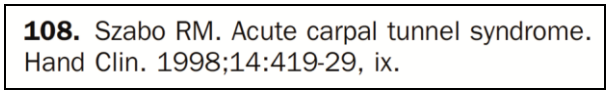


Figure 42: HaCl98_141, citing article

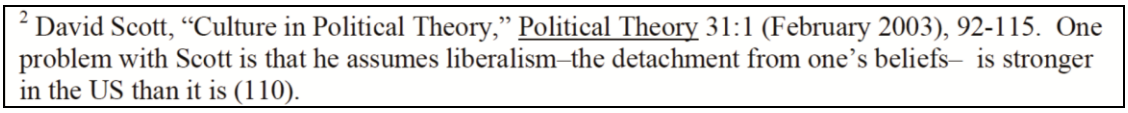


Figure 43: PoTh03_128, citing article

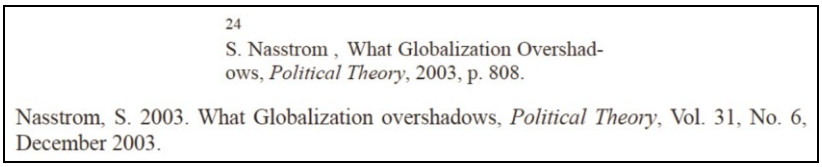


Figure 44: PoTh03_137, citing article

Figure 45: PoTh08_029, citing article

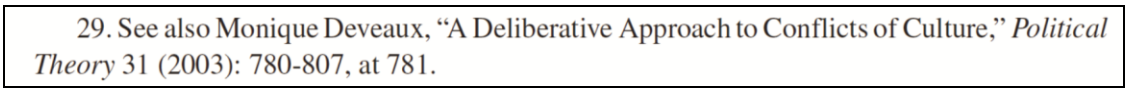


Figure 46: PoTh03_139, citing article

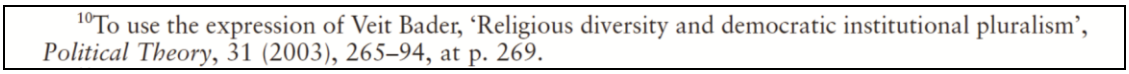


Figure 47: PoTh03_140, citing article

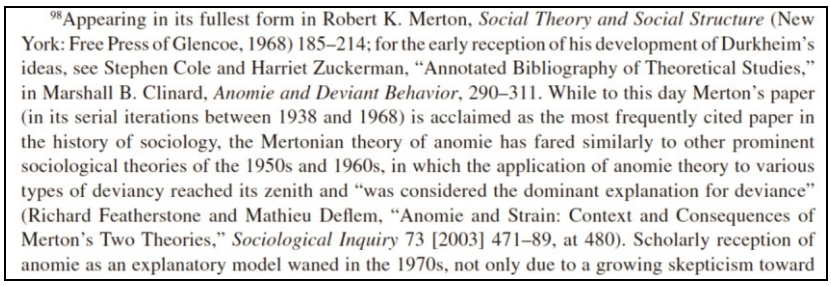


Figure 48: SoIn03_149, citing article

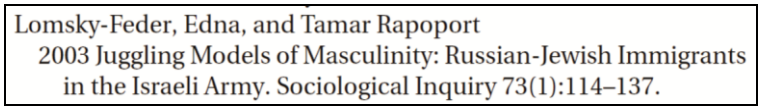


Figure 49: SoIn03_150, citing article

Figure 50: WDMW03_197, citing article

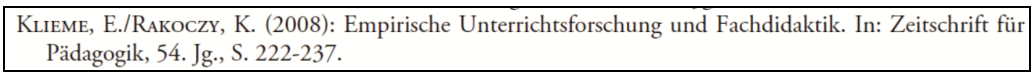


Figure 51: ZPad08_047, citing article

Erklärung über die selbstständige Abfassung meiner Dissertation

Hiermit erkläre ich, Marlies Olensky,

dass ich die vorliegende Dissertation selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe.

Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht.

Die Dissertation wurde bisher in gleicher oder ähnlicher Form keiner anderen Prüfungsbehörde vorgelegt oder veröffentlicht.

Berlin, den 17.10.2014

Unterschrift …….....